Article

Neuron

Causal Inference in the Multisensory Brain

Highlights

- Flexible use of multisensory information follows Bayesian inference
- Sensory fusion emerges earlier than causal inference
- Multisensory inference is represented in the frontal lobe
- Inferior frontal regions guide perception in discrepant environments

Authors

Yinan Cao, Christopher Summerfield, Hame Park, Bruno Lucio Giordano, Christoph Kayser

Correspondence

yinan.cao@psy.ox.ac.uk (Y.C.), christoph.kayser@uni-bielefeld.de (C.K.)

In Brief

Humans combine multisensory information sharing a common cause while avoiding distraction from irrelevant sources. Cao et al. show how this flexible sensory inference is guided by prefrontal cortex and unfolds as sequential neural computations comprising functionally distinct multisensory representations.

Cao et al., 2019, Neuron *102*, 1076–1087
June 5, 2019
https://doi.org/10.1016/j.neuron.2019.03.043

CellPress

CellPress

Neuron
Article

Causal Inference in the Multisensory Brain

Yinan Cao,[1,6,*] Christopher Summerfield,[1] Hame Park,[2] Bruno Lucio Giordano,[3,4,5] and Christoph Kayser[2,5,*]
[1]Department of Experimental Psychology, University of Oxford, Walton Street, Oxford OX2 6AE, UK
[2]Department for Cognitive Neuroscience and Cognitive Interaction Technology-Center of Excellence, Bielefeld University, 33615 Bielefeld, Germany
[3]Institut de Neurosciences de la Timone UMR 7289 Centre National de la Recherche Scientifique and Aix-Marseille Université, Marseille, France
[4]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, UK
[5]Senior author
[6]Lead Contact
*Correspondence: yinan.cao@psy.ox.ac.uk (Y.C.), christoph.kayser@uni-bielefeld.de (C.K.)
https://doi.org/10.1016/j.neuron.2019.03.043

SUMMARY

When combining information across different senses, humans need to flexibly select cues of a common origin while avoiding distraction from irrelevant inputs. The brain could solve this challenge using a hierarchical principle by deriving rapidly a fused sensory estimate for computational expediency and, later and if required, filtering out irrelevant signals based on the inferred sensory cause(s). Analyzing time- and source-resolved human magnetoencephalographic data, we unveil a systematic spatiotemporal cascade of the relevant computations, starting with early segregated unisensory representations, continuing with sensory fusion in parietal-temporal regions, and culminating as causal inference in the frontal lobe. Our results reconcile previous computational accounts of multisensory perception by showing that prefrontal cortex guides flexible integrative behavior based on candidate representations established in sensory and association cortices, thereby framing multisensory integration in the generalized context of adaptive behavior.

INTRODUCTION

We experience the world via multiple sensory modalities. Where information arrives simultaneously in two modalities with differing reliability, the most precise estimates are formed when signals are combined in proportion to their relative reliability. For example, imagine trying to follow a drama on a broken television. If the TV audio is faulty, a viewer should rely more on the picture to follow the narrative, and vice versa. One influential theory suggests that the brains of humans and other animals have evolved to implement this reliability-weighting principle when judging sensory signals (Alais and Burr, 2004; Angelaki et al., 2009; Ernst and Bülthoff, 2004; Raposo et al., 2012). A challenge for the nervous system, however, is that sensory signals should only be fused when they originate from a common source. For instance, if there is a chance that a film is dubbed, combining information about lip movements with prosody will render the dialogue difficult to understand. To meet this challenge, the brain must infer the probability that sensory signals share a common cause (or, to continue the example, that the film was dubbed or not). There is evidence from psychophysics that our brain indeed carries out causal inference (CI) to achieve behavioral flexibility during multisensory integration (Körding et al., 2007; Kayser and Shams, 2015). For example, when localizing auditory and visual signals, we tend to fuse these when they likely originate from nearby sources, but not when they originate from disparate locations, suggesting that the probability of fusion is determined by high-level inference over the probable cause(s) of sensation (De Corte et al., 2018; Körding et al., 2007; Rohe and Noppeney, 2015a, 2015b; Wozny et al., 2010).

Reliability-weighted fusion and CI have complementary costs and benefits. Fusion may allow rapid inference through frugal computations—for example, implemented in feedforward circuits (Ma et al., 2006; Ohshiro et al., 2011, 2017)—and serves as a good rule of thumb for many circumstances that give rise to correlated multimodal signals (Parise and Ernst, 2016). CI permits adaptive behavior but may be slower and more computationally costly, as it requires inference to be carried out over potential states of the world (Kayser and Shams, 2015). We do not understand how the brain arbitrates between the expediency of fusing multimodal signals and the imperative to perform CI in service of optimal perception. One possibility is that the brain hedges its bets by both computing a rapid fused estimate and, later and where required, inferring the likely cause(s) of multimodal signals. This prediction can be evaluated using time-resolved neuroimaging methods, such as magnetoencephalography (MEG), that are equipped to measure how neural signals unfold during a single decision.

Here, combining a multivariate analysis approach to MEG data with computational modeling of behavior, we asked where in the brain and when neural signals predicted by models of sensory fusion and CI emerge during perception. In line with past results, we predicted that fused estimates would be emerging rapidly and in parietal-temporal association cortices (Beauchamp et al., 2004a; Boyle et al., 2017; Calvert, 2001; Macaluso and Driver, 2005). We further reasoned that CI would rely at least in part on the frontal cortex, a structure that is thought to subserve causal

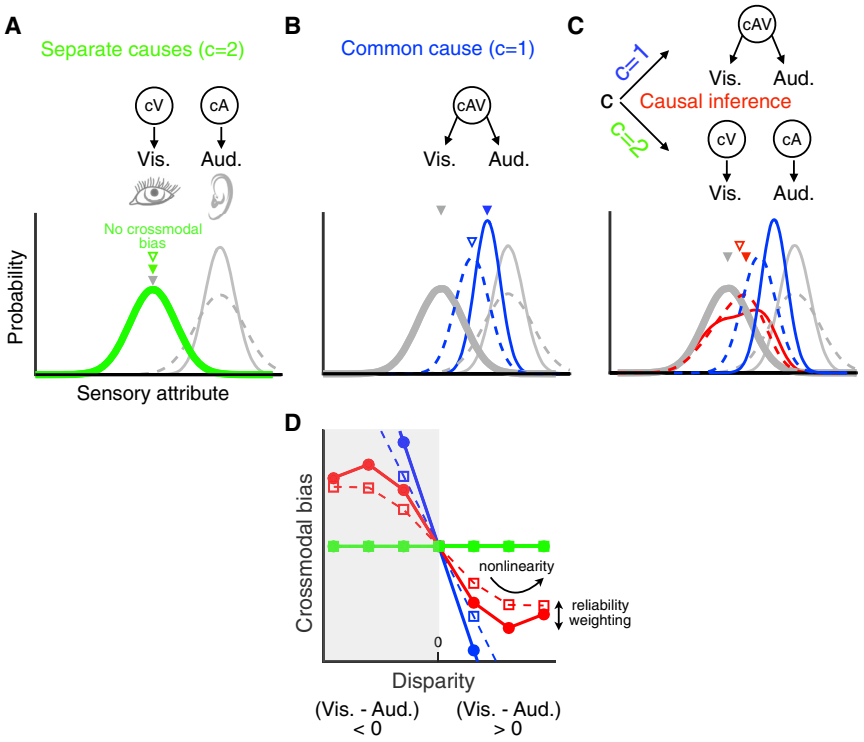

Figure 1. Computational Models

Schematic of different sensory causal structures giving rise to visual and acoustic stimuli. Top of (A) to (C): inferred causality. Bottom: probability distribution of the perceived stimulus feature (e.g., event rate here) and of the sensory estimate derived under different assumptions about the causal structure. Solid/dashed distributions indicate conditions with high/low auditory reliability. Upside-down triangle(s) represent the mean(s) of the distribution(s), with gray denoting task-relevant cue (Vis. in this figure) and colored denoting final estimates under different causal structures (solid versus empty for high versus low reliability of task-irrelevant cue, respectively).

(A) Assuming separate sources for two stimuli (cause [c] = 2) leads to a sensory estimate that reflects the most likely stimulus in the task-relevant modality.

(B) The assumption of a common source leads to the integration of both senses (reliability-weighted fusion).

(C) With causal inference, the two hypotheses about the causal structures (c = 1 or c = 2) are combined probabilistically. The final Bayesian estimate combines the unisensory (task-relevant) and the fused estimates, each weighted by its inferred probability based on an *a priori* integration tendency (STAR Methods).

(D) Each candidate model predicts a unique relationship between crossmodal disparity (distinct visual versus auditory rates are characterized by a large disparity) and bias (deviation of the final estimate from the true attribute). The shaded area corresponds to the example shown in (A) to (C), i.e., visual rate < auditory rate.

reasoning and sensory conflict resolution across a wide range of tasks (Donoso et al., 2014; Giordano et al., 2017; Koechlin and Summerfield, 2007; Noppeney et al., 2010; Tomov et al., 2018).

Using a multisensory rate-categorization paradigm, we show that flexible multisensory behavior is best described by a Bayesian causal inference model. We also find that the neural representations of sensory fusion and inference unfold hierarchically in time and across brain regions. This comprises a cascade from early unisensory encoding in primary sensory cortices to reliability-weighted fusion in parietal-temporal cortices and to CI primarily in the frontal lobe. We find that neural representations within the dorsomedial and ventrolateral PFC are directly predictive of categorical choices and that the vlPFC subserves a particular behavioral benefit of inferring sensory causes to minimize perceptual bias in discrepant crossmodal contexts. Our results reconcile previous rival computational models of multisensory integration (Figure 1) by showing that distinct computational strategies are orchestrated as a temporal sequence and along a parietal-frontal hierarchy. These results also suggest that the neurocomputational mechanisms underlying flexible multisensory perception can be understood in a more general framework of causal reasoning that subserves adaptive behavior in ambiguous environments.

RESULTS

Fifteen human volunteers participated in an audiovisual rate categorization task (four-choice speeded judgement; Figure 2A) while their brain activity was measured using magnetoencephalography (MEG). The stimuli consisted of a temporal sequence of audiovisual pulses (flutter and flicker; duration of the entire sequence was 550 ms) presented at four possible repetition rates (9.1, 12.7, 16.4, or 20 Hz; i.e., number of pulses/s). In separate blocks, participants were instructed to report either the auditory or visual rate as task-relevant information and signaled their response with a button press. Unlike paradigms that focus on sensory fusion by requiring explicit use of both sensory modalities (Ernst and Bülthoff, 2004), our task permits the analysis of individual and flexible strategies in processing multisensory cues based on their task relevance. To quantify how the discrepancy of crossmodal information influences behavior (Körding et al., 2007), we manipulated visual and auditory rates independently (i.e., they could be either congruent or incongruent across trials; Figure 2B). To quantify the reliability-dependent influence of one modality onto another (Angelaki et al., 2009), we varied the signal-to-noise ratio of the acoustic information. The paradigm thus comprised a factorial 4 (visual rates) by 4 (auditory rates) by 2 (auditory reliabilities) by 2 (task relevance) design (see STAR Methods).

Modeling Behavior

We compared the predictions of three classes of models about participants' behavior. Each model encodes probability distributions over sensory signals and incorporates rules that govern how a prior belief about the sensory causal structure is combined with incoming information to judge the event rate in the task-relevant modality (Figure 1). Specifically, we considered (1) a model

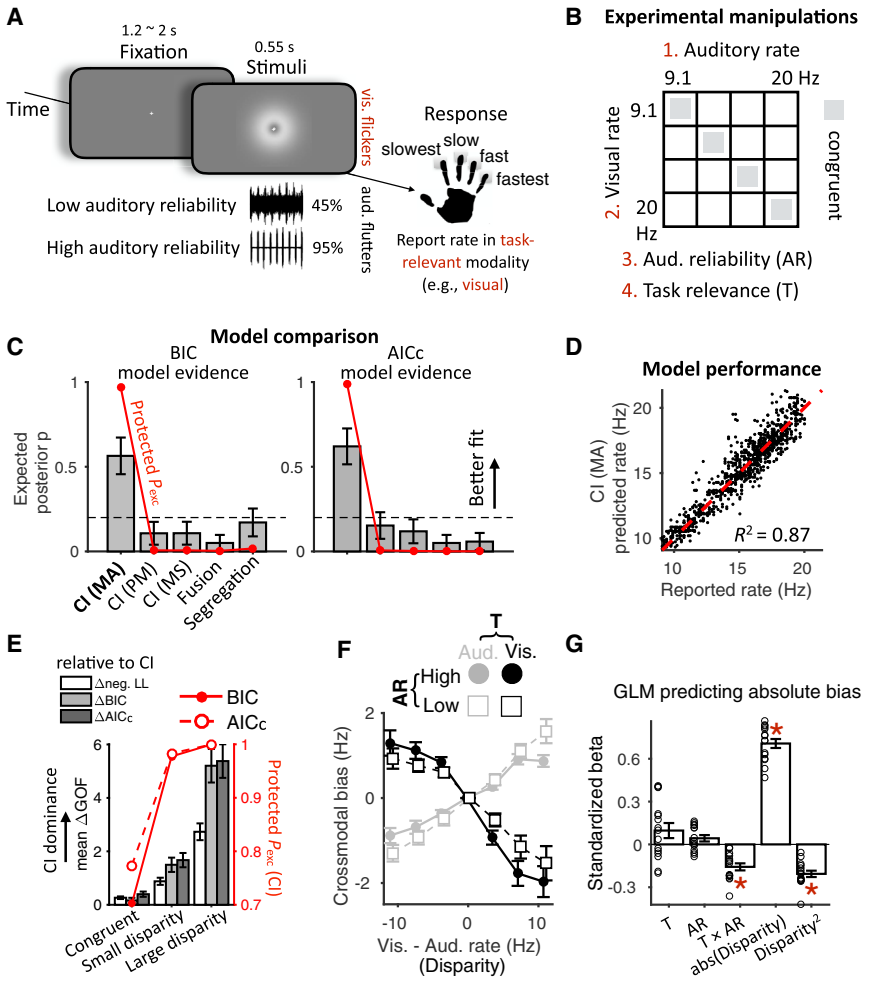

Figure 2. Behavioral Paradigm and Data Analysis

(A) In separate blocks, participants reported either the auditory or visual rate as task-relevant information in a 4-choice speeded categorization task. The panel illustrates the structure of a multisensory trial.

(B) The experimental design featured 4 orthogonal factors: visual and auditory rates, auditory reliability, and task-relevant modality (visual versus auditory).

(C) Model comparison: CI (causal inference), with decision strategy of MA (model averaging), PM (probability matching), or MS (model selection). Fusion: a linear combination of audiovisual information following a belief in a common cause; Segregation: focus on a single modality following a belief in separate causes. Red curves, protected exceedance probability P_{exc} (STAR Methods); gray bars, expected posterior probabilities; error bars, standard deviation (SD) of posterior probabilities; dashed lines, chance level (p = 0.2).

(D) Rate estimates as predicted by the winning CI model in comparison with observed (trial-averaged) rates. Each dot represents an individual condition of one participant. R^2 = generalized coefficient of determination.

(E) Model comparison (CI (MA) versus Fusion versus Segregation) as a function of disparity confirms the leading role of CI, specifically at high disparities. CI dominance is measured as the mean goodness-of-fit (GOF) difference (log-likelihood, BIC or AICc; mean across Fusion - CI and Segregation - CI), and the protected P_{exc} of CI (against Fusion and Segregation, using both BIC and AICc model evidence, same as in C). Congruent: disparity = 0 Hz (16 conditions), small: disparity = 3.64 Hz (24 conditions), large: disparity > 3.64 Hz (24 conditions).

(F) Crossmodal bias, reflecting the disparity-dependent influence of the task-irrelevant cue. Reported rates are grouped by task (T) and auditory reliability (AR). Disparity is signed: visual minus auditory rate.

(G) Standardized regression coefficients quantifying the influence of task (T; visual task minus auditory task), auditory reliability (AR; low minus high), and the linear and quadratic effects of the absolute disparity on the absolute bias (STAR Methods). Red asterisk, significant (2-sided permutation tests; family-wise-error corrected $p \leq 0.05$); error bars, ± 1 SEM (n = 15); circles, participant-specific betas. See also Tables S1 and S2 and Figure S1.

of "sensory segregation," (2) a model for reliability-weighted "sensory fusion" (Ernst and Bülthoff, 2004), and (3) Bayesian models of multisensory "causal inference" (Körding et al., 2007; Wozny et al., 2010).

These models make distinct predictions about how the perceived event rate varies with experimental manipulations (crossmodal disparity, i.e., the difference between auditory and visual rates; cue reliability; and task relevance). One key behavioral variable is the level of *crossmodal bias*, i.e., the extent to which judgements about the relevant modality are biased by the irrelevant modality, and how this bias varies with disparity. The segregation model proposes that sensory estimates are fully independent and predicts no crossmodal bias. The fusion model instead predicts a bias that grows linearly with disparity because relevant and irrelevant sensory signals are fused irrespective of their congruency. This model does, however, predict that bias will scale with the reliability of individual cues. Finally, the inference model allows for an additional inference about sensory causality, i.e., that observers allow for some signals to be fused and some to be segregated and that fusion is more likely for signals that are similar in rate (Körding et al., 2007; Rohe and Noppeney, 2015b; Wozny et al., 2010). This inference model predicts that the bias increases with disparity and relative cue reliabilities (Rohe and Noppeney, 2015a), but critically, it predicts that the growth rate of bias should diminish for highly discrepant information that is unlikely to originate from a common source, i.e., reflecting a nonlinear dependency of bias on disparity, in contrast to the fusion model predicting a linear dependency. Hence, among these candidates, only the inference model reflects the behavioral flexibility to exploit multisensory information when of benefit, and to otherwise avoid distraction from cues that likely have an independent origin.

Multisensory Judgments Follow Bayesian Causal Inference

We determined which candidate model best accounts for participants' behavior. We maximized each model's log-likelihood of

explaining individual participant's responses across all conditions and derived the best-fitting model parameters (Table S1 for model comparison; Table S2 for parameter estimates). A Bayesian causal inference model formulated with a free probabilistic belief of common cause (p_c) and with a model-averaging decision strategy, "CI (MA)," explained the data better than models not incorporating the inference of latent cause(s) (i.e., segregation and reliability-weighted fusion, Figure 2C; group-level Bayesian (BIC) and corrected Akaike Information Criterion (AICc) relative to CI (MA) $\geq$ 468 and 547, respectively). Model averaging here refers to a decision function that averages the fused and the task-relevant segregated sensory estimates, each weighted by their inferred probabilities. The CI (MA) also fared best in random-effects model comparison (Figure 2C; protected exceedance probability $P_{exc} = 0.967$ and 0.989 using BIC and AICc model evidence, respectively). Other variants of the CI model with alternative decision strategies ("probability matching" and "model selection," see STAR Methods) provided worse fit than the CI model with model-averaging strategy (protected $P_{exc} < 0.0062$). Across participants, the average coefficient of determination R^2 of this best CI model was 0.87 (SEM = 0.0078; Figure 2D). Quantifying CI model performance as a function of disparity further emphasized that participants' behavior was best explained by CI for discrepant contexts (Figure 2E).

Context-Dependent Cue Weighting

We further examined why CI outperforms the other models in describing the behavioral responses by using an alternative analysis. Specifically, we quantified crossmodal bias, defined as the deviation of participants' response from the actual task-relevant rate (Figure 2F), and used a general linear model (GLM) to predict how the magnitude (i.e., absolute value) of this bias depended on the contextual factors: task, reliability, and their interaction, as well as disparity (Figure 2G; all effects were assessed using maximum-statistics permutation controlling for multiple comparisons, family-wise error FWE = 0.05). Importantly, we included an effect of squared disparity in this model to capture whether the bias scales nonlinearly with disparity, as predicted by CI, or simply follows a linear dependency, as predicted by sensory fusion. A reliability-weighted cue combination is captured by the interaction between task and reliability rather than by the main effect of reliability. This is because reliability was manipulated only for the acoustic signal, which, under reliability-based cue weighting, would result in different biases for the two tasks. Indeed, this GLM revealed no main effects of task ($t(14) = 1.84$, mean $\beta = 0.097$, SEM = 0.053) and auditory reliability ($t(14) = 1.91$, mean $\beta = 0.043$, SEM = 0.022) but a significant interaction between task relevance and auditory reliability ($t(14) = -6.36$, mean $\beta = -0.16$, SEM = 0.025). Lastly, the GLM revealed a significantly negative effect of squared disparity ($t(14) = -9.28$, mean $\beta = -0.21$, SEM = 0.022), confirming a reduction of the bias growth rate for larger disparities (i.e., nonlinear scaling) as suggested by CI.

As expected, reaction times (RTs) also varied systematically with the experimental manipulations (Figures S1A and S1B). In particular, participants' responses were generally slower (yet with RT increasing nonlinearly) for larger disparity, suggesting an additional effort required to make judgements facing disparate multisensory evidence. In addition, the RTs in congruent multisensory trials were significantly faster than those in unisensory trials with the same rate and task ($t(14) = -2.6$, p = 0.021; 2-sided paired-sample t test), confirming a general facilitation of responses by multisensory congruency.

MEG Data Reveal a Spatiotemporal Hierarchy of Multisensory Representations

Next, we investigated when and where the brain represents sensory information in a manner as predicted by each computational model. Specifically, we asked whether neural representations of different candidate computations emerge simultaneously or sequentially and possibly within the same or across distinct brain regions. We adopted a multivariate approach using cross-validated representational similarity analysis (RSA, Walther et al., 2016). RSA assesses neural representations by quantifying the statistical association between two dissimilarity matrices: (1) an MEG representational dissimilarity matrix (RDM) that quantifies the pairwise dissimilarities of brain activity in response to different experimental conditions and (2) a model RDM that quantifies the hypothesized computational nature of brain representations. The model RDMs were constructed based on the rate estimates predicted by each model fitted to the behavioral responses of each participant (Figure 3A). Given the condition-wise differences in reaction times (Figure S1), we carried out separate RSAs by aligning the data to stimulus onset and to trial-by-trial response time.

The results of the stimulus-locked RSA revealed a systematic and gradual progression of neural representations from segregated unisensory representations to reliability-weighted fusion, to CI across cortical space and time (Figure 3A; Table 1). We quantified the selective representation of each model (e.g., CI) that is not already explained by the representations of other models (fusion and segregation) using semi-partial correlation (s.p.r). The earliest MEG activations significantly reflecting one of the candidate computations ("RSA effects," hereafter) were those pertaining to segregated unisensory estimates starting around 100 ms after the stimulus onset. These were localized within the respective sensory cortices (bilateral calcarine cortex starting from ~100 ms for segregated visual representations; auditory cortex starting from ~140 ms for segregated auditory representations). Subsequently, the MEG activity began to reflect sensory representations formed by reliability-weighted fusion, with significant clusters emerging around 180 ms to 260 ms in the left superior temporal gyrus and later the precuneus and superior parietal lobule, the ventral posterior cingulate, and the posterior superior temporal gyrus. Finally, MEG activity reflecting representations as predicted by CI emerged around 620 ms in dorso- and ventrolateral prefrontal cortices (the left inferior frontal cortex, in particular), frontopolar and insular cortices, and the middle-posterior cingulate cortex.

The results of the response-locked RSA also suggested a difference between parietal and frontal regions in reflecting fusion and CI. This revealed an RSA effect of fusion within the parietal-temporal lobe (right precuneus and IPS around −220 ms to −140 ms prior to response onset; Figure 3A; Table 1) but CI within the frontal lobe (bilateral IFG and frontal pole around −220 ms to −140 ms; bilateral superior frontal gyri around

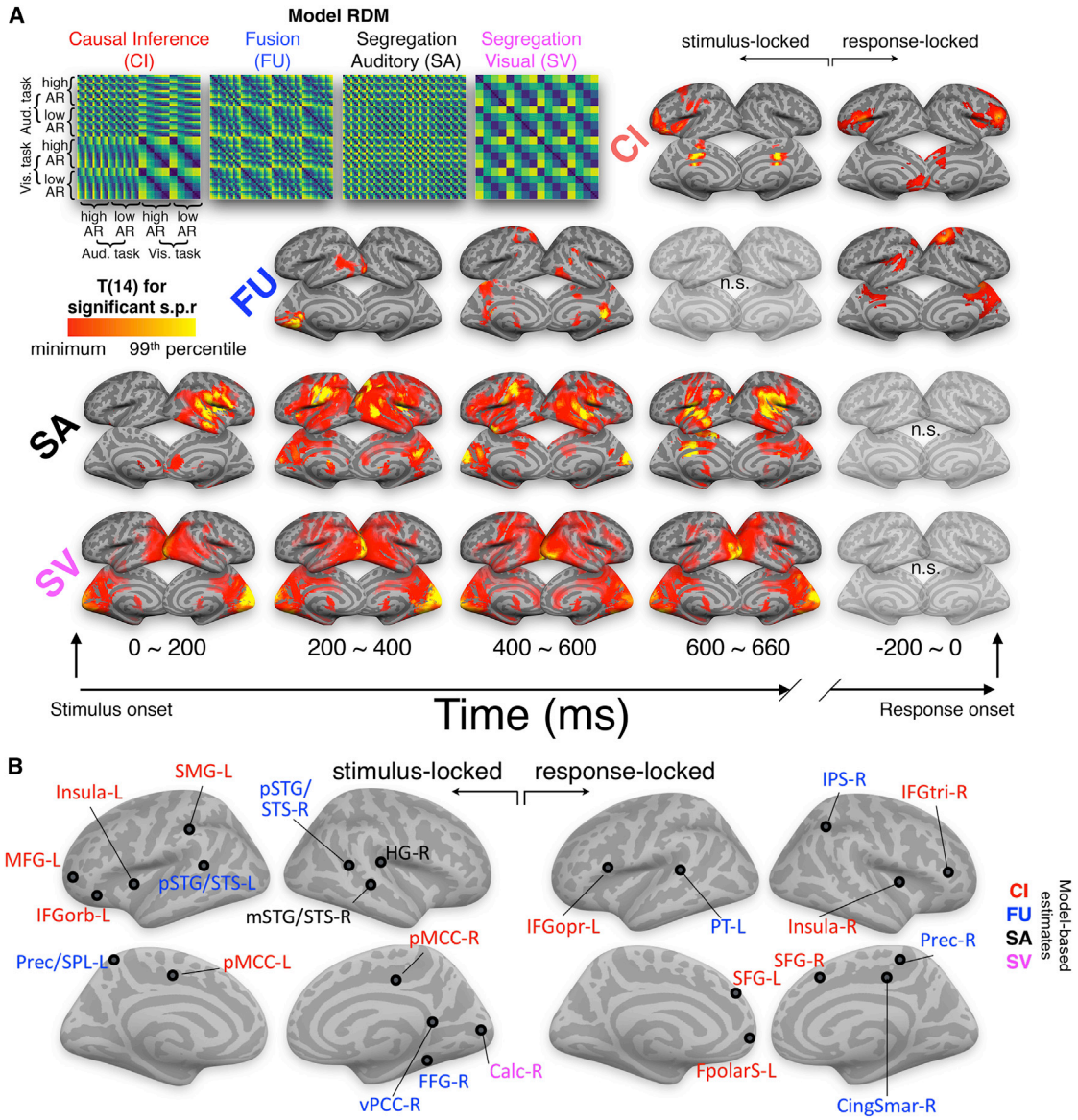

Figure 3. Spatiotemporal Evolution of Multisensory Representations Revealed by Model-Based Representational Similarity Analysis (RSA)

(A) Upper left: group-averaged model representational dissimilarity matrices (RDMs) visualizing the ranked distance between model-predictions of condition-wise event rates. The other figures display the cortical surface projection onto a FreeSurfer template of the T-maps of group-level significant MEG activity uniquely explained by each model (semi-partial correlation, s.p.r) in stimulus- and response-locked data. All effects were significant at $p \leq 0.05$ FWE-corrected (STAR Methods). The maps show peak T-value of each voxel across time within the respective epoch for visualization purpose. Light empty surfaces denote non-significance (n.s.).

(B) Anatomical locations of local and global peaks of RSA effects, color coded by the respective significant model (see also Table 1 and Figure S4). See also Figures S2 and S3.

−180 ms to −100 ms). Encoding of segregation models was no longer found in this response-locked analysis.

To further substantiate the observation that the representations of fusion precede those of CI, we explicitly characterized the temporal sequence of the peak RSA effects for fusion and CI. We extracted the latency of the peak T statistics of each RSA effect (Figure S3) and derived a latency contrast for each pair of the respective ROIs (latency of fusion minus latency of CI). Across the 30 pairs of stimulus-locked ROIs, the peak RSA effects of fusion emerged significantly earlier than those of CI (two-sided one-sample t test against zero: $t(29) = -7.19$, $p < 10^{-7}$, 95% bias-corrected and accelerated [BCa] bootstrap confidence interval of the mean [−0.21, −0.12] s; Figures S3C and S3D). This result was also confirmed for the response-locked ROIs ($t(23) = -3.24$, $p = 0.004$, 95% BCa bootstrap confidence interval of the mean [−0.023, −0.006] s; Figures S3E and S3F).

Table 1. Selective Model Encoding in MEG Source Activity

Analysis: Model	Anatomical Label	Latency	T (14)	s.p.r (SEM)	MNI Coordinates		
Stimulus locked: Segregation Visual (SV)	Calc-R*	140 ms	18.96	0.126 (0.007)	14	−84	8
Stimulus locked: Segregation Auditory (SA)	mSTG/STS-R	180 ms	3.08	0.034 (0.011)	48	−32	−2
	HG-R*	220 ms	4.34	0.026 (0.006)	34	−28	12
Stimulus locked: Fusion (FU)	pSTG/STS-L	220 ms	2.94	0.021(0.007)	−60	−50	8
	Prec/SPL-L	500 ms	3.42	0.02 (0.006)	−10	−56	62
	pSTG/STS-R	500 ms	3.33	0.013 (0.004)	46	−42	16
	vPCC-R*	540 ms	4.29	0.02 (0.005)	10	−46	6
	FFG-R	540 ms	2.96	0.02 (0.007)	34	−46	−20
Stimulus locked: Causal Inference (CI)	pMCC-L*	620 ms	4.26	0.014 (0.003)	−10	−14	44
	pMCC-R	620 ms	4.03	0.018 (0.004)	10	−22	36
	IFGorb-L	620 ms	3.44	0.018 (0.005)	−42	24	−8
	Insula-L	620 ms	3.39	0.024 (0.007)	−38	0	2
	MFG-L	620 ms	3.25	0.017 (0.005)	−42	48	2
	SMG-L	620 ms	2.73	0.012 (0.004)	−56	−32	40
Response locked: FU	IPS-R*	−180 ms	5.14	0.036 (0.007)	34	−52	44
	Prec-R	−180 ms	5.05	0.037 (0.007)	4	−38	54
	CingSmar-R	−180 ms	4.39	0.018 (0.004)	14	−28	36
	PT-L	−180 ms	3.22	0.025 (0.008)	−50	−38	8
Response locked: CI	IFGopr-L*	−180 ms	4.96	0.022 (0.005)	−42	10	8
	FpolarS-L	−180 ms	4.55	0.023 (0.005)	−14	60	−12
	IFGtri-R	−180 ms	4.54	0.018 (0.004)	34	32	12
	Insula-R	−180 ms	3.65	0.015 (0.004)	32	0	16
	SFG-R	−140 ms	3.45	0.019 (0.005)	14	20	48
	SFG-L	−140 ms	3.01	0.01 (0.003)	−4	46	36

The table lists global (*) and local peaks of RSA effects. MNI, Montreal Neurological Institute. Anatomical labels are based on Automated Anatomical Labeling (AAL) and Destrieux atlases. Latency, center of RSA window. s.p.r, group-averaged semi-partial rank correlation between MEG and model RDM. SEM, standard error of the mean. All reported effects are significant at $p \leq 0.05$ FWE corrected (see STAR Methods). L/R, left/right hemisphere. See also Figure S2. HG, Heschl's gyrus; mSTG/STS, middle superior temporal gyrus/sulcus; pSTG/STS, posterior superior temporal gyrus/sulcus; Calc, calcarine; Prec, precuneus; SPL, superior parietal lobule; STGlat, lateral superior temporal gyrus; FFG, fusiform gyrus; vPCC, posterior-ventral cingulate gyrus; SMG, supramarginal gyrus; pMCC, middle-posterior cingulate gyrus and sulcus; IFGorb, inferior frontal gyrus (pars orbitalis); MFG, middle frontal gyrus. IPS, intraparietal sulcus; PT, planum temporale; CingSmar, marginal branch of the cingulate sulcus; IFGopr, inferior frontal gyrus (pars opercularis); FpolarS, transverse frontopolar sulcus; IFGtri, inferior frontal gyrus (pars triangularis); SFG, superior frontal gyrus.

Distinct Markers of Sensory Integration in Posterior and Anterior Brain Regions

We characterized the computational properties of the neural representations revealed by the RSA in further detail. Note that these neural representations are defined both by their location (ROI) and by latency of the respective RSA effect (Table 1). First, we quantified the dependency of the crossmodal bias captured by these representations on the two key features of context-dependent integration: (1) the interaction between task relevance and sensory reliability, and (2) the quadratic effect of disparity. In line with the predictions of fusion and inference (CI) models, we expected fusion regions to exhibit a positive linear dependency of bias on disparity and an interaction between task and reliability. By contrast, CI regions should exhibit a nonlinear relationship between bias and disparity. For this analysis, we focused on computation-diagnostic neural representations that potentially contain these key markers (see STAR Methods).

As expected, a significant task × reliability interaction in shaping bias emerged in all parietal and posterior-temporal fusion ROIs ($t(14) \leq -4.84$, FWE-corrected $p < 0.012$). We also found this interaction in the primary auditory cortex (HG-R, $t(14) = -8.34$, $p < 0.0001$) and some ROIs exhibiting RSA CI effects (SMG and insula; $t(14) \leq -5.45$, $p < 0.0036$; Figure 4A left). By contrast, the nonlinear dependency of bias on disparity characteristically distinguished CI from fusion ROIs, as only frontal CI regions (e.g., Fpolar and IFGtri) exhibited such well-pronounced negative dependency on squared disparity ($t(14) \leq -5.74$, $p < 0.0022$ for the 9 significant CI ROIs in Figure 4A right).

These results suggest a posterior-to-anterior gradient of the influence of multisensory disparity (Figure 4B). To support this, we correlated the GLM coefficients of the squared disparity with the MNI posterior-anterior coordinates of the CI ROIs. For both stimulus- and response-locked ROIs, this correlation was significant ($t(14) \leq -6.62$, $p < 10^{-5}$), with across-participants mean (SEM) of −0.34 (0.05) and −0.68 (0.06), respectively. In line with an asymmetry between the primary visual and auditory cortex (Rohe and Noppeney, 2016), we also observed a significant negative quadratic influence of disparity in occipital

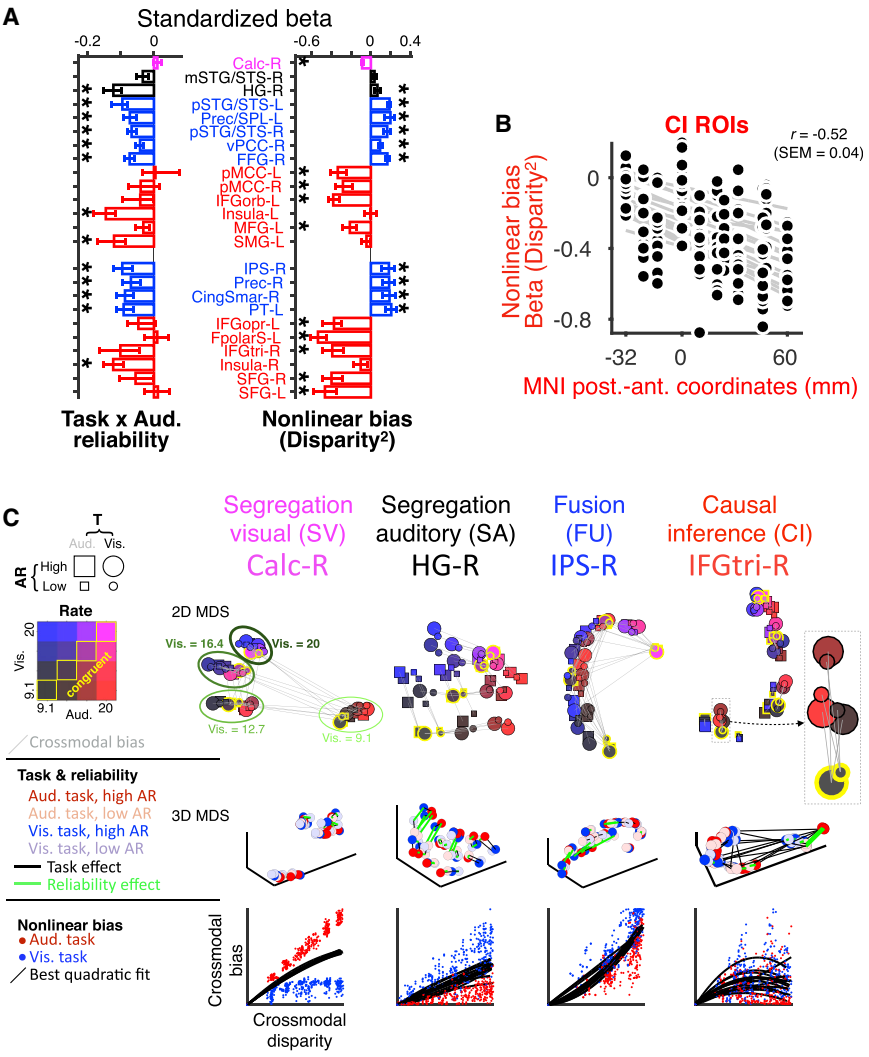

Figure 4. Functional and Anatomical Characterization of the Neural Representations of Interest

(A) General linear modeling (GLM) of the influence of the task × reliability interaction (signature of reliability-weighted fusion) and the quadratic disparity (signature of causal inference) on cross-modal bias. GLM was applied to the computation-diagnostic RDMs from each ROI (see STAR Methods). Bar height and error bars, mean and 95% bootstrap confidence interval of the standardized GLM coefficients across participants (cf. Figure S4 for all ROIs; asterisk = $p \leq 0.05$ FWE-corrected; STAR Methods).

(B) Negative correlation between anatomical MNI posterior-anterior coordinates and GLM betas for quadratic disparity across CI ROIs. Dots, individual participants for each ROI; gray lines, linear fit for each participant.

(C) Inter-individual difference multidimensional scaling (MDS) was used to visualize the representational geometry of neural RDMs. Experimental conditions that are closer in MDS space evoke similar computation-diagnostic MEG responses (STAR Methods). For simplicity, we focus on one region for each model (see Figure S5 for all ROIs). The two-dimensional (2D) MDS models emphasize the computational relevance of the two main dimensions: auditory (horizontal) and visual rate (vertical); gray lines indicate the neural distances reflecting crossmodal bias. Shaded green ellipses highlight the 4 primary clusters separating conditions associated with the 4 visual rates in the 2D MDS of Calc-R. Zoomed cluster in 2D MDS of IFGtri-R illustrates disparity-dependent nonlinear bias: representations at higher crossmodal disparity (red circles) being pulled toward the task-relevant rate (yellow contours). The three-dimensional (3D) MDS models emphasize the effects of task relevance and auditory reliability (color coded). The scatterplots visualize the neurally encoded disparity against crossmodal bias. See also Figures S4 and S5.

(Calc-R in Figure 4A, $t(14) = -17.6$, $p = 10^{-4}$; also see Figure 4C), but not temporal, regions.

In a second analysis, we modeled the representational geometries of computation-diagnostic representations using multi-dimensional scaling (MDS). This projects the neural representations onto a few axes that capture the dimensions along which the candidate representations differ (de Leeuw and Mair, 2009; Ashby et al., 1994). Exemplar ROIs for each computational model are shown in Figure 4C (see Figure S5 for all ROIs). In early visual cortex (Calc), the representation varied primarily with visual rate, irrespective of task and reliability. By contrast, in early auditory regions (HG), the representations were modulated by auditory reliability and by task and reflected an influence from the visual modality, resulting in a seemingly bimodal representation (2D MDS). In ROIs reflecting fusion (e.g., IPS-R), the representational geometry collapsed the rates across modalities and disparities into a single dimension (leading to large biases) but changed with reliability (green lines in 3D MDS). Finally, in ROIs reflecting CI (e.g., IFGtri-R), the representational geometry varied with all three factors (nonlinear scaling with disparity in 2D MDS; reliability and task in 3D MDS) and hence exhibited the highest computational flexibility among all ROIs.

Complementary Evidence for Sensory Fusion in Parietal-Temporal Regions and Causal Inference in Frontal Regions

To further corroborate the parieto-frontal gradient in the neural representations of sensory fusion (FU) and CI, we implemented a bootstrap-based comparison of each model-predicted sensory representation in explaining the local pattern of MEG activity based on group-level exceedance probabilities (P_{exc}).

ROIs with fusion or CI RSA effects exhibited high P_{exc} for the respective models, in particular in the response-locked data ($P_{exc} > 0.88$ among all 4 fusion ROIs, with the highest P_{exc} (FU) in IPS-R = 0.997; $P_{exc} > 0.81$ among all CI ROIs, with the highest P_{exc}(CI) around FpolarS-L = 0.979; Figure 5A). In the stimulus-locked data, the ROIs with CI RSA effects also exhibited high P_{exc} for CI, and the parietal ROIs with fusion RSA

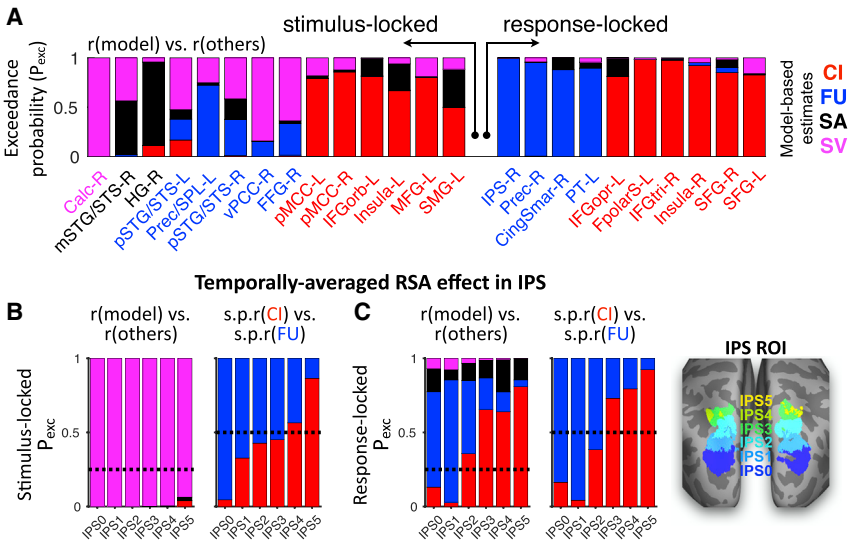

Figure 5. Model Comparison Based on Exceedance Probabilities

Exceedance probabilities (P_{exc}) index the belief that a region encodes a given model more likely than alternative models across participants.

(A) Model P_{exc} for each of the ROIs derived from the RSA (cf. Figure 3).

(B and C) Model P_{exc} within the intraparietal sulcus (IPS) for comparison with Rohe and Noppeney (2015b), for stimulus-locked (B) and response-locked data (C). Left panels: P_{exc} comparing each of the four models with all of the alternative models derived independently for each IPS ROI using time-averaged RSA statistics (encoding coefficient r, Fisher Z scale). Right panels: P_{exc} comparing time-averaged selective encoding (s.p.r, Fisher Z scale) of causal inference (CI) and fusion (FU). Dashed lines, chance level. IPS ROIs were based on a probabilistic atlas (Wang et al., 2015).

effects exhibited high P_{exc} for fusion (Prec/SPL-L, P_{exc}(FU) = 0.72). Some occipital-temporal regions contain a graded organization suggesting also a contribution of segregated sensory representations apart from fusion. The superior temporal lobe comprised not only regions representing fusion (bilateral pSTG/STS) but more "bimodal" regions (mSTG/STS) that maintain the individual unisensory representations.

These results confirm a parieto-frontal gradient whereby the respective regions predominantly encode the sensory estimates based on FU and CI. This is seemingly at odds with a previous study describing a gradient from FU to CI locally within the IPS (Rohe and Noppeney, 2015b; P_{exc} approach to fMRI data). To make the present data directly comparable to this previous study, we quantified model P_{exc} within anatomically defined IPS sub-regions (Wang et al., 2015). This confirmed a gradient along the posterior-anterior axis of the IPS, with the more posterior regions being dominated by FU and the more anterior regions dominated by CI (Figures 5B and 5C). Hence, while our data also support a graded representation of CI in the parietal lobe, the whole-brain analysis here emphasizes a wider network that chiefly involves the frontal cortex.

Prefrontal Cortex Drives Flexible Behavior in Conflicting Multisensory Environments

Lastly, we asked which of the candidate ROIs is directly predictive of behavior. The preceding results reveal that multiple regions reflect distinct multisensory computations. Yet, for a given ROI, a significant model-encoding RSA effect by itself does not imply a driving role for the full flexibility of participants' behavior (Kriegeskorte and Douglas, 2018). Hence, to identify in which candidate ROIs the MEG activity was directly predictive of participants' responses, we implemented an RSA to assess the association between the MEG RDMs and the behavioral RDM (the pairwise absolute distance between the trial-averaged behavioral responses in different conditions). We focused on the ROIs identified in the response-aligned data to account for differences in RT between conditions. As above, we considered each ROI at its specific latencies of model encoding (see Table 1). The CI ROIs within the lateral PFC (IFG and SFG) and insula exhibited significant neuro-behavioral correlations ($t(14) \geq 3.01$, FWE-corrected $p \leq 0.036$; Figure 6). An additional contrast between conditions with large and small crossmodal disparities revealed that activity within the ventrolateral PFC (IFGtri-R) particularly predicted behavioral responses when the two modalities were highly conflicting ($t(14) = 2.94$, $p = 0.043$).

DISCUSSION

We characterized the computational strategies in the service of flexible multisensory integration and identified their time-resolved underpinnings in source-localized MEG activity. Our behavioral data suggest that when faced with trial-varying reliability and discrepancy of crossmodal information, humans arbitrate between integration and segregation in a manner captured by Bayesian causal inference. At the neural level, we unveil that the distinct computations required for flexible multisensory perception (segregation, fusion, and CI) coexist, but each dominates at different times and in distinct regions. Consistent with previous reports, these results show that the initially segregated unisensory signals are fused in temporal and parietal lobes. However, this fused information gives way to more flexible representations formed under multisensory CI in the frontal cortex, which drive behavior in volatile environments with discrepant multimodal cues.

Temporal Hierarchy of Multisensory Computations

Previous studies have posed reliability-weighted fusion and CI as rival accounts of multisensory perception and sought to obtain empirical evidence that favors one over the other (Acerbi et al., 2018; Körding et al., 2007; Magnotti and Beauchamp, 2017; Parise et al., 2012; Roach et al., 2006; Rohe and Noppeney, 2015a; Wozny et al., 2010). In the spirit of this endeavor, we found that human behavior during a rate categorization task was consistently captured by a CI model. This fared best in quantitative model comparison and was the only model able to describe the reduction in crossmodal bias that emerged

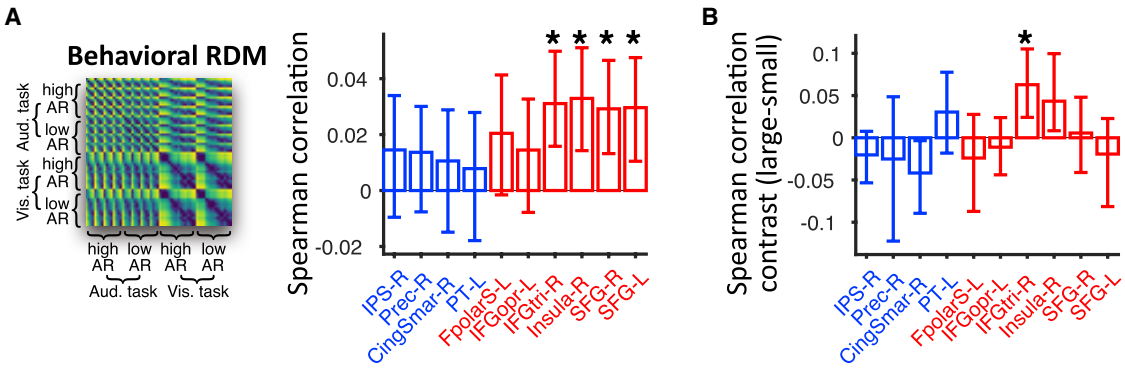

Figure 6. Contribution of Frontal Regions to Flexible Behavior

(A) RSA was applied to assess the neural representation of behavioral reports (behavioral RDM; left shown, average across participants). Bar height and error bars, mean and 95% bootstrap confidence intervals.

(B) Disparity-modulated RSA effect of behavioral report was assessed by contrasting brain-behavior correlations between conditions with small and large disparities (two-sided permutation test; small disparity = 3.64 Hz, large disparity > 3.64 Hz). Asterisk = $p \leq 0.05$ FWE-corrected (see STAR Methods).

when sensory cues were most disparate. Critically, the neural data challenge the dichotomy between fusion and CI as separate accounts of multisensory perception. Rather, they support the notion that perception is a hierarchical process relying on the explicit representations of distinct multisensory computations orchestrated over several brain regions (Rohe and Noppeney, 2015b; Rohe et al., 2019; Kayser and Shams, 2015). Our results suggest that representations as predicted by each model coexist and unveil the functional hierarchy of the underlying computations in distinct regions and over different timescales.

We observed a systematic temporal sequence whereby the neural underpinnings of sensory segregation and reliability-weighted fusion were succeeded by those of CI. This cascade suggests a specific computational scheme for CI at the systems level: CI effectively relies on a weighted combination of sensory estimates predicted by fusion and segregated signals from *a posteriori* estimates, with the relative contribution depending on the inferred level of crossmodal disparity. Thus, one may interpret fusion as a component that explicitly feeds into the computations for inference (Körding et al., 2007; Rohe and Noppeney, 2015a, 2015b; Wozny et al., 2010). Consistent with the early emergence of fusion, previous evidence suggests that multisensory integration starts at the level of early sensory cortices (Kayser and Logothetis, 2007; Lakatos et al., 2007; Lewis and Noppeney, 2010), and neuroimaging data support physiological correlates of reliability-weighted fusion starting around 120 ms post-stimulus onset (Aller and Noppeney, 2019; Boyle et al., 2017). Along this line, behavioral studies also suggested that fusion may be a rather automatic process. For example, crossmodal biases tend to be stronger when participants respond faster or after acquiring only little sensory evidence (Noppeney et al., 2010). By contrast, CI requires additional processing time as it capitalizes on evaluating the degree of sensory discrepancy, maintaining beliefs over latent causes, and possibly exploring distinct decision strategies. Indeed, the conscious segregation of multimodal cues that usually tend to be fused requires additional time and effort (Gau and Noppeney, 2016). In line with these reports, and based on a principled assessment of candidate representations in spatiotemporally resolved brain activity, we demonstrate a systematic emergence of fused sensory representations within parietal-temporal regions before those predicted by CI in frontal regions.

Graded Context-Dependent Computations along Parietal-Frontal Pathway

Behavior adapts to contextual modulations, such as trial-varying reliability of individual senses or changes in the congruency of sensory information (Angelaki et al., 2009; Ma and Pouget, 2008). Previous modeling studies have formalized two hallmarks of context-dependent integration: (1) giving more credence to the more reliable modality (Alais and Burr, 2004; Ernst and Bülthoff, 2004) and (2) refraining from distraction by irrelevant information from an apparently distinct causal origin (Körding et al., 2007; Roach et al., 2006). The observers in our study exhibited both hallmarks as shown by their perceptual biases. Resolving the sensory conflict and inferring the sensory causal structure require time, and this was also visible in the nonlinear dependency of RTs on disparity. Importantly, the observers' brain activities directly reflected both hallmarks of flexible multisensory computation. Our complementary analysis approaches to the MEG data, relying on either selective model encoding or group-level model exceedance probabilities, corroborated a parieto-frontal gradient reflecting a progression from the neural representations adapting to sensory reliability to representations adapting to disparity. While these results resonate with a previous study suggesting the coexistence of distinct context-dependent multisensory computations (Rohe and Noppeney, 2015b), our findings elaborate this perspective of computational gradient along a wider parietal-frontal network that supports CI in more anterior regions. This view also fits with a general understanding of the functional differences between parietal and frontal regions, whereby both accumulate sensory information, but parietal cortex potentially encodes more of the sensory aspect, whereas frontal regions underlie the context-dependent

transformation of evidence into choice (Erlich et al., 2015; Hanks et al., 2015).

Across the brain, we observed multiple and functionally heterogeneous multisensory representations on which flexible inference can capitalize (Bizley et al., 2016). First, we found a weak but noticeable auditory bias in V1, confirming multisensory interactions in early sensory cortex (Kayser and Logothetis, 2007; Lakatos et al., 2007). Second, within the temporal lobe, we revealed the coexistence of bimodal (in the absence of integration) and truly fused representations, in line with a topographical organization of uni- and multisensory representations (Beauchamp et al., 2004a). Third, we found that the neural representation of fusion was stronger in parietal than in temporal regions (Rohe and Noppeney, 2016), which may pertain to the abstract nature of the rate stimuli (Chafee, 2013; Raposo et al., 2014). Yet, these regions did not exhibit the characteristic down-weighting of task-irrelevant information at larger disparities indicative of inference. Rather, parietal activity appeared to scale with the task-irrelevant information, suggesting that these regions may automatically project evidence to one dimension (Suzuki and Gottlieb, 2013; Ganguli et al., 2008; Luyckx et al., 2019).

By contrast, we found that the frontal cortex has a privileged role in reflecting both hallmarks of context-dependent integration but features subtle representational variations across regions. Notably, the nonlinear influence of disparity increased systematically along the posterior-to-anterior axis. While representations within more posterior regions (e.g., the insula) appear to emphasize the reliability-based weighting, representations in anterior regions mostly depend on disparity (e.g., frontopolar cortex and inferior frontal gyrus). Such heightened sensitivity of more anterior regions to the causal structure resonates with the general hierarchical organization of the PFC, whereby anterior regions are involved in representing abstract rules (Badre and D'Esposito, 2009). By combining the localization of specific candidate representations in MEG activity with an analysis probing their choice-predictability, our results further demonstrate the behavioral relevance of frontal representations. Here the more prominent role of the dorsomedial and ventrolateral, rather than the anterior, PFC in driving behavior fits with the role of these regions in inferring probable causes of observed contingencies for retrieving goal-relevant behavioral strategies (Donoso et al., 2014).

The Frontal Lobe in Multisensory and Domain-General Inference

A key aspect of behavioral flexibility is the ability to perceive the environments through multiple senses and to exploit the most appropriate modalities for the task at hand. Given the role of the frontal cortex in subserving general reasoning and adaptive behavior (such as inferring the reliability of different decision strategies), it is perhaps not surprising that we found the key driver of flexible multisensory behavior in the frontal lobe (Donoso et al., 2014; Koechlin and Summerfield, 2007; Tomov et al., 2018). Yet, previous studies promoted divergent views on the neural basis of multisensory perception. In fact, many studies have emphasized the role of superior temporal and parietal cortices in sensory integration (Beauchamp et al., 2004b; Calvert, 2001). In part, this may have arisen from the specific search for bimodal representations in some studies and for sensory fusion in others, or the aim to identify the earliest convergence of multisensory signals (Kayser and Logothetis, 2007). However, neuroanatomical evidence has long implied the prefrontal cortex as a convergence zone for multisensory information (Jones and Powell, 1970), with the ventrolateral prefrontal cortex receiving projections from auditory and visual cortices and association areas implementing sensory fusion (Romanski, 2012). In the context of multisensory integration, the PFC has been highlighted as a domain-general structure responsible for evaluating sensory discordance (Adam and Noppeney, 2010; Calvert, 2001; Hein et al., 2007; Noppeney et al., 2010) and has been implied in forming beliefs over inferential states, based on expectations or prior experience, in face of sensory uncertainty (Gau and Noppeney, 2016; Kayser and Kayser, 2018; Noppeney et al., 2010).

Yet, the specific computations underlying frontal multisensory representations and their role in driving behavior remained elusive. Our results reconcile the previous literature and suggest that the frontal cortex implements a flexible strategy capitalizing on distinct computational solutions (fusion versus segregation) when processing volatile multisensory information. Thereby perception effectively amplifies the behavioral significance either of segregated representations established early on in a trial and maintained within sensory-specific regions, or of fused representations formed within temporal and parietal association regions later on. These results help to unify previous studies on general adaptive behavior with those elucidating the function of frontal regions in multisensory integration. In fact, prefrontal cortex seems to support a domain-general mechanism for selecting among multiple candidate strategies for solving a problem at hand, when handling both multisensory and other information. As such, the role of the PFC is unlikely about merging sensory information alone, but about arbitrating between competing strategies of how the most appropriate sensory representation should be formed to guide behavior.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Task design and stimuli
  - Experimental procedure and stimulus presentation
  - Neuroimaging data acquisition
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Analysis of behavioral data - Crossmodal bias
  - Analysis of behavioral data – Modeling
  - Analysis of behavioral data – Model fitting
  - Analysis of Behavioral Data – Model Comparison
  - Analysis of behavioral data – Sensory noise function
  - Preprocessing of MEG Data
  - Reconstruction of MEG sources

- Representational similarity analysis
- Analysis of the neural representations in regions of interest (ROIs)
- Analysis of cerebral representation of categorization behavior

- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.neuron.2019.03.043.

ACKNOWLEDGMENTS

We thank Robin A.A. Ince for support during initial development of MEG analysis pipeline and Andreas Jarvstad, Jan Balaguer, and Valentin Wyart for helpful discussions. Y.C. was funded by the University of Oxford Clarendon Fellowship, the Department of Experimental Psychology, and a Brasenose-Kwai Cheong Studentship. This project was funded by the European Research Council (to C.K. ERC-2014-CoG; grant No 646657), UK's Biotechnology and Biological Sciences Research Council (grant BB/M009742/1 to B.L.G.), and a Human Brain Project award to C.S. B.L.G. was further supported by ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and Excellence Initiative of Aix-Marseille University (A*MIDEX).

AUTHOR CONTRIBUTIONS

Conceptualization: Y.C., C.K.; Methodology: Y.C., B.L.G., C.K.; Software: Y.C., B.L.G.; Validation: Y.C., B.L.G., C.K.; Formal Analysis: Y.C., B.L.G., C.K.; Investigation: Y.C., H.P., B.L.G.; Resources: Y.C., B.L.G., C.K.; Data Curation: Y.C., B.L.G.; Writing – Original Draft: Y.C., B.L.G., C.K., C.S.; Writing – Review & Editing: Y.C., C.S., H.P., B.L.G., C.K.; Visualization: Y.C., B.L.G.; Supervision: C.K., B.L.G., C.S.; Project Administration: B.L.G., C.K.; Funding Acquisition: C.K., B.L.G., C.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 20, 2018
Revised: February 18, 2019
Accepted: March 27, 2019
Published: April 29, 2019

REFERENCES

Acerbi, L., and Ma, W.J. (2017). Practical Bayesian optimization for model fitting with Bayesian Adaptive Direct Search. In Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 1836–1846.

Acerbi, L., Dokka, K., Angelaki, D.E., and Ma, W.J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. PLoS Comput. Biol. *14*, e1006110.

Adam, R., and Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. Neuroimage *52*, 1592–1602.

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. *14*, 257–262.

Aller, M., and Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian Causal Inference. PLoS Biol. *17*, e3000210.

Angelaki, D.E., Gu, Y., and DeAngelis, G.C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation. Curr. Opin. Neurobiol. *19*, 452–458.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. Neuroimage *38*, 95–113.

Ashby, F.G., Maddox, W.T., and Lee, W.W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. Psychol. Sci. *5*, 144–151.

Badre, D., and D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? Nat. Rev. Neurosci. *10*, 659–669.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., and Martin, A. (2004a). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat. Neurosci. *7*, 1190–1192.

Beauchamp, M.S., Lee, K.E., Argall, B.D., and Martin, A. (2004b). Integration of auditory and visual information about objects in superior temporal sulcus. Neuron *41*, 809–823.

Bizley, J.K., Jones, G.P., and Town, S.M. (2016). Where are multisensory signals combined for perceptual decision-making? Curr. Opin. Neurobiol. *40*, 31–37.

Boyle, S.C., Kayser, S.J., and Kayser, C. (2017). Neural correlates of multisensory reliability and perceptual weights emerge at early latencies during audio-visual integration. Eur. J. Neurosci. *46*, 2565–2577.

Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

Calvert, G.A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex *11*, 1110–1123.

Cavanaugh, J.E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. Stat. Probab. Lett. *33*, 201–208.

Chafee, M.V. (2013). A scalar neural code for categories in parietal cortex: representing cognitive variables as "more" or "less". Neuron *77*, 7–9.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. PLoS Comput. Biol. *10*, e1003441.

De Corte, B.J., Della Valle, R.R., and Matell, M.S. (2018). Recalibrating timing behavior via expected covariance between temporal cues. eLife *7*, e38790.

de Leeuw, J., and Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. J. Stat. Softw. *31*, 1–30.

Donoso, M., Collins, A.G., and Koechlin, E. (2014). Human cognition. Foundations of human reasoning in the prefrontal cortex. Science *344*, 1481–1486.

Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap (CRC press).

Erlich, J.C., Brunton, B.W., Duan, C.A., Hanks, T.D., and Brody, C.D. (2015). Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. eLife *4*, e05457.

Ernst, M.O., and Bülthoff, H.H. (2004). Merging the senses into a robust percept. Trends Cogn. Sci. *8*, 162–169.

Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

Ganguli, S., Bisley, J.W., Roitman, J.D., Shadlen, M.N., Goldberg, M.E., and Miller, K.D. (2008). One-dimensional dynamics of attention and decision making in LIP. Neuron *58*, 15–25.

Gau, R., and Noppeney, U. (2016). How prior expectations shape multisensory perception. Neuroimage *124* (Pt A), 876–886.

Giordano, B.L., Ince, R.A.A., Gross, J., Schyns, P.G., Panzeri, S., and Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. eLife *6*, e24763.

Giordano, B.L., Whiting, C., Kriegeskorte, N., Kotz, S.A., Belin, P., and Gross, J. (2018). From categories to dimensions: spatio-temporal dynamics of the cerebral representations of emotion in voice. bioRxiv. Published online February 15, 2018. https://doi.org/10.1101/265843.

Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature *520*, 220–223.

Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., and Cichy, R.M. (2018). The representational dynamics of task and object processing in humans. eLife *7*, e32816.

Hein, G., Doehrmann, O., Müller, N.G., Kaiser, J., Muckli, L., and Naumer, M.J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J. Neurosci. *27*, 7881–7887.

Hipp, J.F., and Siegel, M. (2013). Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in EEG. Front. Hum. Neurosci. *7*, 338.

Jones, E.G., and Powell, T.P.S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. Brain *93*, 793–820.

Kayser, S.J., and Kayser, C. (2018). Trial by trial dependencies in multisensory perception and their correlates in dynamic brain activity. Sci. Rep. *8*, 3742.

Kayser, C., and Logothetis, N.K. (2007). Do early sensory cortices integrate cross-modal information? Brain Struct. Funct. *212*, 121–132.

Kayser, C., and Shams, L. (2015). Multisensory causal inference in the brain. PLoS Biol. *13*, e1002075.

Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. Trends Cogn. Sci. *11*, 229–235.

Körding, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., and Shams, L. (2007). Causal inference in multisensory perception. PLoS ONE *2*, e943.

Kriegeskorte, N., and Douglas, P.K. (2018). Cognitive computational neuroscience. Nat. Neurosci. *21*, 1148–1160.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

Lakatos, P., Chen, C.M., O'Connell, M.N., Mills, A., and Schroeder, C.E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron *53*, 279–292.

Lewis, R., and Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. J. Neurosci. *30*, 12329–12339.

Luyckx, F., Nili, H., Spitzer, B., and Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. eLife *8*, e42816.

Ma, W.J., and Pouget, A. (2008). Linking neurons to behavior in multisensory perception: a computational review. Brain Res. *1242*, 4–12.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Macaluso, E., and Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci. *28*, 264–271.

MacKay, D.J. (2003). Information theory, inference and learning algorithms (Cambridge, UK: Cambridge University Press).

Magnotti, J.F., and Beauchamp, M.S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. PLoS Comput. Biol. *13*, e1005229.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. Biometrika *78*, 691–692.

Nieder, A., and Miller, E.K. (2003). Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex. Neuron *37*, 149–157.

Noppeney, U., Ostwald, D., and Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. J. Neurosci. *30*, 7434–7446.

Ohshiro, T., Angelaki, D.E., and DeAngelis, G.C. (2011). A normalization model of multisensory integration. Nat. Neurosci. *14*, 775–782.

Ohshiro, T., Angelaki, D.E., and DeAngelis, G.C. (2017). A neural signature of divisive normalization at the level of multisensory integration in primate cortex. Neuron *95*, 399–411.e8.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. *2011*, 156869.

Parise, C.V., and Ernst, M.O. (2016). Correlation detection as a general mechanism for multisensory integration. Nat. Commun. *7*, 11543.

Parise, C.V., Spence, C., and Ernst, M.O. (2012). When correlation implies causation in multisensory integration. Curr. Biol. *22*, 46–49.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. *10*, 437–442.

Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., and Wu, X. (2009). Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. IEEE Trans. Audio Speech Lang. Process. *17*, 1124–1132.

Raposo, D., Sheppard, J.P., Schrater, P.R., and Churchland, A.K. (2012). Multisensory decision-making in rats and humans. J. Neurosci. *32*, 3726–3735.

Raposo, D., Kaufman, M.T., and Churchland, A.K. (2014). A category-free neural population supports evolving demands during decision-making. Nat. Neurosci. *17*, 1784–1792.

Rigoux, L., Stephan, K.E., Friston, K.J., and Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. Neuroimage *84*, 971–985.

Roach, N.W., Heron, J., and McGraw, P.V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. Proc. Biol. Sci. *273*, 2159–2168.

Rohe, T., and Noppeney, U. (2015a). Sensory reliability shapes perceptual inference via two mechanisms. J. Vis. *15*, 22.

Rohe, T., and Noppeney, U. (2015b). Cortical hierarchies perform Bayesian causal inference in multisensory perception. PLoS Biol. *13*, e1002073.

Rohe, T., and Noppeney, U. (2016). Distinct computational principles govern multisensory integration in primary sensory and association cortices. Curr. Biol. *26*, 509–514.

Rohe, T., Ehlis, A., and Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian inference in multisensory perception. Nat. Commun. Published December 21, 2018. https://doi.org/10.1101/504845.

Romanski, L.M. (2012). Convergence of auditory, visual, and somatosensory information in ventral prefrontal cortex. In The neural bases of multisensory processes, *Chapter 33*, M.M. Murray and M.T. Wallace, eds. (Boca Raton, FL: CRC Press/Taylor & Francis).

Seibold, D.R., and McPhee, R.D. (1979). Commonality analysis: A method for decomposing explained variance in multiple regression analyses. Hum. Commun. Res. *5*, 355–365.

Suzuki, M., and Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. Nat. Neurosci. *16*, 98–104.

Tomov, M.S., Dorfman, H.M., and Gershman, S.J. (2018). Neural computations underlying causal structure learning. J. Neurosci. *38*, 7143–7157.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage *137*, 188–200.

Wang, L., Mruczek, R.E., Arcaro, M.J., and Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. Cereb. Cortex *25*, 3911–3931.

Wozny, D.R., Beierholm, U.R., and Shams, L. (2010). Probability matching as a computational strategy used in perception. PLoS Comput. Biol. *6*, e1000871.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Behavioral data and MEG RDMs	This paper	https://github.com/YinanCao/Causal-Inference
Software and Algorithms		
MATLAB R2015B	MathWorks	https://www.mathworks.com/
Psychtoolbox-3	Brainard, 1997; Pelli, 1997	http://psychtoolbox.org/
R version 3.4.0	R Development Core Team	https://www.r-project.org/
SPM12	Wellcome Trust	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
FreeSurfer	Fischl, 2012	https://surfer.nmr.mgh.harvard.edu/
PKU & IOA HRTF database	Qu et al., 2009	http://www.cis.pku.edu.cn/auditory/Staff/Dr.Qu.files/Qu-HRTF-Database.html
Fieldtrip	Oostenveld et al., 2011	http://www.fieldtriptoolbox.org/
BADS	Acerbi and Ma, 2017	https://github.com/lacerbi/bads
SMACOF	de Leeuw and Mair, 2009	https://cran.r-project.org/web/packages/smacof/index.html
MTIMESX	James Tursa	https://www.mathworks.com/matlabcentral/fileexchange/25977-mtimesx-fast-matrix-multiply-with-multi-dimensional-support
VBA toolbox	Daunizeau et al., 2014	https://mbb-team.github.io/VBA-toolbox/
Custom code (analyses)	This paper	https://github.com/YinanCao/Causal-Inference

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for sources and reagents should be directed to and will be fulfilled by the Lead Contact, Yinan Cao (yinan.cao@psy.ox.ac.uk), Department of Experimental Psychology, University of Oxford.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sixteen right-handed adults (8 females; aged 19 to 35 years, mean = 24.9, SD = 4.8) participated in this study. All reported normal hearing and vision, were briefed on the nature and goal of this study, and received financial compensation for their participation. The study was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committee (College of Science and Engineering, University of Glasgow). Written informed consent was obtained from all participants prior to the study. The head movement of one male participant during MEG acquisition was excessive (see Preprocessing of MEG Data) thus the data were not included in the analysis. All results are reported for an N of 15.

METHOD DETAILS

Task design and stimuli

Participants categorized the temporal rate of stochastic sequences of brief visual and auditory pulses (flicker/flutter; see Roach et al., 2006 for conceptually similar stimuli), presented either in unisensory or multisensory conditions. In each block participants were instructed to report either the auditory or visual rate as the task-relevant information. They were asked to respond as accurately and as quickly as possible and to think of the visual and auditory pulses as originating either from independent sources or from a common generative process. Such a design allows for an analysis of participants' flexible strategies concerning when and how to use multisensory cues for a judgement based on their task relevance. The pulse-trains (full duration = 550 ms) consisted of sequential 16.7-ms visual flickers (Gaussian annulus) and auditory flutters (amplitude-modulated white noise) with congruent or incongruent rates (each being one of four possible levels: 9.1, 12.7, 16.4 or 20 Hz, i.e., 5, 7, 9 or 11 events presented in 550 ms). The first/last auditory flutter was always onset-synchronized with the first/last visual flicker, while intermediate pulses were temporally jittered. The inter-event intervals (IEIs; time from one event's offset to the onset of a subsequent event) of a perfectly periodic event sequence were perturbed using a zero-mean Gaussian jitter with a standard deviation equal to the 10% of the periodic-sequence IEI. The jittered visual events

were further rounded to the nearest possible video flips and had a minimal IEI of 2 visual frames to avoid flicker fusion (event-onset-asynchrony $\geq$ 50 ms). The jitter was independent for the two modalities rendering the auditory and visual pulses temporally asynchronous regardless of the overall congruency in auditory and visual rates.

The visual stimulus was a difference of two concentric Gaussians (standard deviation of 3° and 1.5° respectively; resulting in a Gaussian annulus that made it possible to present a constant central fixation cross – white, diameter = 0.4°– throughout the entire experiment). The auditory stimuli were noise bursts generated by modulating (square wave) the amplitude of a 550-ms white noise with 3-ms onset-/offset cosine ramp. We manipulated the modulation depth – either 95% or 45% of the peak amplitude – resulting in two levels of auditory reliability. These sound stimuli were then convolved with a head-related transfer function (elevation = 0°, azimuth = 0°; PKU-IOA HRTF database, Qu et al., 2009) to promote perceived co-localization of the sound with the visual stimuli projected to an external screen. The brightness of the visual flicker was jittered on a trial-by-trial basis to de-correlate the visual contrast from sound intensity. Across trials, the auditory and visual rates, auditory reliability and the task-relevant modality were manipulated factorially in a 4 (visual rates) by 4 (auditory rates) by 2 (auditory reliabilities) by 2 (task relevance) design, resulting in 64 multisensory conditions.

Experimental procedure and stimulus presentation

Each run of MEG data collection (5 min) contained all conditions (64 multisensory and 12 unisensory). Within each run the auditory and visual tasks were separated into two task-specific blocks starting with on-screen instructions about the relevant modality for upcoming trials. Unisensory and multisensory conditions were interleaved within each task-specific block. The auditory block contained 32 multisensory trials plus 8 auditory trials (4 rates × 2 reliabilities), whereas the visual block consisted of 32 multisensory trials plus 4 visual trials (4 rates). The order of auditory and visual blocks was counterbalanced across runs. Prior to the experiment participants were passively exposed to 64 trials of unisensory stimuli (visual and auditory separately) with repeatedly increasing rates (slowest to fastest, repeated 8 times) and were instructed to memorize the 4 rate categories for use in the following main task. No feedback on correctness or speed was provided during the main experiment.

Stimulus presentation was controlled in MATLAB (Mathworks, Inc, Natick, USA) using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Visual stimuli were projected through a DLP projector (Panasonic D7700) onto a non-translucent screen at 1280 × 720 pixels at 60 fps covering a field of view of 25° × 19°. The viewing distance was 180 cm. Acoustic flutters (sampling rate = 48 kHz; depth = 16 bits; 80 dB SPL root-mean-square value, measured with a Brüel & Kjær Type 2205 sound-level meter, A-weighting) were presented through Etymotic ER-30 insert tubephones. We used inverse filtering methods to eliminate the spectral distortion of the sound stimuli induced by the frequency response of the tubephone system (Giordano et al., 2018).

Neuroimaging data acquisition

MEG data were acquired using a 248-magnetometers whole-head MEG system (MAGNES 3600 WH, 4-D Neuroimaging) at a sampling rate of 1017.25 Hz with participants seated upright. The positions of five coils marking fiducial landmarks on the head of the participants were acquired at the beginning and end of each run. Runs associated with excessive head movements, MEG noise or containing reference-channel jumps were discarded. For each participant we selected the acceptable 22 noise- and artifact-free runs with the smallest average head movement. Across these, the maximum change in head position was on average 3.9 mm (SD = 1.12 mm). To facilitate source analysis, we acquired for each participant the digitized head shape and the locations of the five coils, as well as a whole-brain, high-resolution, structural T1-weighted MP-RAGE image (192 sagittal slices, 256 × 256 matrix size, 1 mm^3 voxel size; Siemens 3T Trio scanner; 32-channel head coil).

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of behavioral data - Crossmodal bias

We quantified the magnitude of the crossmodal bias as the absolute deviation of the reported rate from the task-relevant rate in each trial (Figure 2F and 2G). To correct for a potential response tendency toward intermediate rates, we adjusted the task-relevant rate using the mean reported rate derived from congruent trials using an established method (Rohe and Noppeney, 2015a). Specifically, we replaced the task-relevant physical rate with the trial-averaged perceived rate in the respective congruent conditions. This adjustment was carried out independently for each participant, and for each task and reliability level. We used a general linear model (GLM) to predict the magnitude of the bias using the experimental factors and specific interactions of these (Figure 2G):

$$\begin{aligned}\text{abs(crossmodal bias)} \sim{}& \beta_0 + \beta_1\,\text{T} + \beta_2\,\text{AR} + \beta_3(\text{T} \times \text{AR}) + \beta_4 \\ &\text{abs(Disparity)} + \beta_5\,\text{Disparity}^2\end{aligned} \qquad \text{(Equation 1)}$$

where T = task relevance, AR = auditory reliability. GLM predictors were Z-scored to obtain comparable effect sizes. Significance testing relied on a permutation procedure shuffling the condition labels and we corrected for multiple comparisons using maximum-statistics (10,000 permutations; family-wise error FWE = 0.05 across GLM effects). The results of this analysis were also validated in a separate analysis that considered a more complex GLM with additional higher-order effects (Figure S1D), or a GLM with additional cubic effect of disparity (resulting for all participants in an increase of BIC relative to the original GLM; average BIC increase = 2.9; SEM = 0.31). We emphasize that the goal of this GLM was not to provide the best possible mathematical description for the bias

curve, but, most importantly, to identify the behavior with a key prediction of causal inference: the deviation from a linear dependency of bias on disparity as suggested by the fusion model.

Analysis of behavioral data – Modeling

We fitted three classes of candidate models as computational accounts for participants' responses. First, we considered a unisensory segregation model, which predicts responses for each condition as the unisensory rate perceived in the task-relevant modality. Second, we fitted an established maximum-likelihood model for reliability-weighted sensory fusion (Ernst and Bülthoff, 2004). This predicts the response as a linear combination of the task-relevant and task-irrelevant sensory cues, each weighted in proportion to their relative reliability (inverse variance). Importantly, this model reflects a mandatory integration of the two cues, regardless of their discrepancy. This model is known to account for multisensory behavior in conditions where sensory discrepancies are small and both senses are required for a judgement (Ernst and Bülthoff, 2004). Third, we considered a class of Bayesian models of causal inference describing the arbitration between perceptual strategies in variable multisensory environments (Körding et al., 2007; Wozny et al., 2010; Rohe and Noppeney, 2015a). These models incorporate a probabilistic belief about the causal relation between the sensory inputs (Figure 1) and use this to arbitrate (in a statistical sense) between integrating and segregating the evidence from the different senses.

For all models, the sensory estimate is about event rate, *r*. We modeled the respective unisensory estimates using Gaussian sensory likelihoods: $p(r_a|s_a) = \mathrm{N}(s_a,\ \sigma_a^2)$ and $p(r_v|s_v) = \mathrm{N}(s_v,\ \sigma_v^2)$, with *s* denoting the physical rate of stimulus, and $\sigma_a^2(\sigma_v^2)$ the variance of the sensory likelihood distribution of unisensory auditory (visual) estimates, respectively. The sensory likelihood thus represents the probability of experiencing sensation *r* as a result of stimulus *s* occurring in the environment. To capture response bias toward intermediate rates, we included a Gaussian central prior, $p(s) = \mathrm{N}(\mu_p, \sigma_p^2)$, with variable mean μ_p and variance σ_p^2.

Here, we briefly introduce the key aspects of each model (details can be found in: Ernst and Bülthoff, 2004; Körding et al., 2007; Wozny et al., 2010). First, we assumed that the reported rates are conditioned upon causal structure: either two independent causes ($c = 2$, Figure 1 left, Equation 2) or a single common cause ($c = 1$, Figure 1 middle, Equation 3). Then the posterior probability of auditory or visual cue given a specific causal structure *c* and sensory observation *r*, is given by Bayes' rule:

$$p(s_a|r_a, c=2) = \frac{p(r_a|s_a)p(s)}{p(r_a)}$$

$$p(s_v|r_v, c=2) = \frac{p(r_v|s_v)p(s)}{p(r_v)} \qquad \text{(Equation 2)}$$

$$p(s\,|\,r_v, r_a, c=1) = \frac{p(r_a|s)p(r_v|s)p(s)}{p(r_v, r_a)} \qquad \text{(Equation 3)}$$

Under the assumption of Gaussian distributions, the maximum-a-posteriori estimates are given by Equation 4 for segregation and by Equation 5 for fusion:

$$\hat{S}_{a,\,c=2} = \frac{r_a/\sigma_a^2 + \mu_p/\sigma_p^2}{1/\sigma_a^2 + 1/\sigma_p^2},\ \hat{S}_{v,\,c=2} = \frac{r_v/\sigma_v^2 + \mu_p/\sigma_p^2}{1/\sigma_v^2 + 1/\sigma_p^2} \qquad \text{(Equation 4)}$$

$$\hat{S}_{a,\,c=1} = \hat{S}_{v,\,c=1} = \frac{r_a/\sigma_a^2 + r_v/\sigma_v^2 + \mu_p/\sigma_p^2}{1/\sigma_a^2 + 1/\sigma_v^2 + 1/\sigma_p^2} \qquad \text{(Equation 5)}$$

For the causal inference models, we modeled an ideal Bayesian observer who uses an inferred belief about the multisensory causality ($c = 1$ for common cause, $c = 2$ for independent causes) to estimate the rate. This belief (Equations 8 and 9) is probabilistically determined by combining the sensory likelihoods with an integration tendency p_c, that is, an *a priori* belief in a common cause across the full experimental settings. Given the Gaussian assumption, the sensory likelihoods $p(r_v, r_a|c)$ of the visual and auditory estimates (r_v and r_a) under each causality assumption are:

$$p(r_v,\ r_a\,|\,c=1) = \frac{1}{2\pi\sqrt{\sigma_v^2\sigma_a^2 + \sigma_v^2\sigma_p^2 + \sigma_a^2\sigma_p^2}}\, e^{-\frac{1}{2}\left[\frac{(r_v-r_a)^2\sigma_p^2 + (r_v-\mu_p)^2\sigma_a^2 + (r_a-\mu_p)^2\sigma_v^2}{\sigma_v^2\sigma_a^2 + \sigma_v^2\sigma_p^2 + \sigma_a^2\sigma_p^2}\right]} \qquad \text{(Equation 6)}$$

$$p(r_v,\ r_a \mid c=2) = \frac{1}{2\pi\sqrt{\left(\sigma_v^2+\sigma_p^2\right)\left(\sigma_a^2+\sigma_p^2\right)}} e^{-\frac{1}{2}\left[\frac{\left(r_v-\mu_p\right)^2}{\left(\sigma_v^2+\sigma_p^2\right)}+\frac{\left(r_a-\mu_p\right)^2}{\left(\sigma_a^2+\sigma_p^2\right)}\right]^2} \quad \text{(Equation 7)}$$

Given these sensory likelihoods and integration tendency, the posterior belief about the causal structure can be inferred using Bayes' rule:

$$p(c=1 \mid r_v,\ r_a) = \frac{p(r_v,\ r_a \mid c=1)p_c}{p(r_v,\ r_a \mid c=1)p_c + p(r_v,\ r_a \mid c=2)(1-p_c)} \quad \text{(Equation 8)}$$

$$p(c=2 \mid r_v,\ r_a) = 1 - p(c=1 \mid r_v,\ r_a) \quad \text{(Equation 9)}$$

Decision strategies of causal inference. We considered three decision strategies when modeling how the inferred causal structure (Equations 8 and 9) is used to obtain a rate estimate: Model averaging (MA; Körding et al., 2007), probability matching (PM) and model selection (MS; Wozny et al., 2010). MA (Equation 10) minimizes the mean squared error of the final estimate by combining the estimates derived from segregation and fusion (in Equations 4 and 5), each weighted by the inferred posterior probability over the respective causal structure:

$$\hat{S}_a = p(c=1 \mid r_v, r_a)\hat{S}_{a,\ c=1} + p(c=2 \mid r_v, r_a)\hat{S}_{a,\ c=2}$$

$$\hat{S}_v = p(c=1 \mid r_v, r_a)\hat{S}_{v,\ c=1} + p(c=2 \mid r_v, r_a)\hat{S}_{v,\ c=2} \quad \text{(Equation 10)}$$

With PM as decision strategy (Equation 11) an observer aims to arbitrate between the fusion and segregation using a probabilistic rule, with the relative probability of each outcome matching the inferred probability of each causal scenario. This strategy was modeled using a stochastic selection criterion, γ, which was sampled uniformly in [0, 1] and independently on each trial:

$$\hat{S}_a = \hat{S}_{a,\ c=1} \text{ if } p(c=1 \mid r_v, r_a) > \gamma,\ \hat{S}_a = \hat{S}_{a,\ c=2} \text{ if } p(c=1 \mid r_v, r_a) \leq \gamma$$

$$\hat{S}_v = \hat{S}_{v,\ c=1} \text{ if } p(c=1 \mid r_v, r_a) > \gamma, \hat{S}_v = \hat{S}_{v,\ c=2} \text{ if } p(c=1 \mid r_v, r_a) \leq \gamma \quad \text{(Equation 11)}$$

With MS as decision strategy an observer selects fusion (segregation) as long as the inferred posterior probability of the respective causal structure "c = 1" ("c = 2") exceeds 0.5. Practically, this corresponds to fixing γ in Equation 11 at 0.5.

We also considered a model in which the observer consistently holds a neutral belief about a common cause (p_c = 0.5) thus only uses sensory likelihood to determine the causal structure. We termed this a "likelihood" model because it assumes that the posterior over a causal structure can be fully determined by the ratio of sensory likelihood itself (Equation 12). For this model the posterior probability of a common cause is given by:

$$p(c=1 \mid r_v, r_a) = \frac{1}{1 + p(r_v, r_a \mid c=2)/p(r_v, r_a \mid c=1)} \quad \text{(Equation 12)}$$

Analysis of behavioral data – Model fitting

To estimate the best-fitting model parameters, for each participant we implemented an optimization search that maximized the log-likelihood of each model given the participant's data. Suppose the counts of the four choices in a given condition are N_i, with $i = \{1, 2, 3, 4\}$. Then the log-likelihood of a model that predicts the response probabilities p_i associated with each of the 4 choices is given by:

$$LL(model \mid \{N_i\}) = \ln C + \sum_{i=1}^{4} N_i \ln p_i \quad \text{(Equation 13)}$$

where C denotes the multinomial coefficient, i.e., $C = (\sum_{i=1}^{4} N_i)!/(\prod_{i=1}^{4} N_i!)$. Maximizing Equation 13 is equivalent to maximizing $\sum_{i=1}^{4} N_i \ln p_i$ because $\ln C$ is a constant. However, $\ln C$ must be estimated for calculating the generalized coefficient of determination (see Model Comparison below). In practice, C can be written as a product of a series of binomial coefficients:

$$C = \binom{N_1}{N_1}\binom{N_1+N_2}{N_2}\binom{N_1+N_2+N_3}{N_3}\binom{N_1+N_2+N_3+N_4}{N_4} \quad \text{(Equation 14)}$$

For large N_i, the logarithm of the factorial function (Equation 14) can be approximated using Stirling's formula (MacKay, 2003):

$$\ln\frac{a!}{b!(a-b)!}\approx b\ln\frac{a}{b}+(a-b)\ln\frac{a}{a-b} \quad \text{(Equation 15)}$$

Parameter estimation relied on the Bayesian Adaptive Direct Search (BADS; Acerbi and Ma, 2017) maximizing the sum of log-likelihoods across all 64 multisensory conditions. To avoid local optima, for each model we generated a grid of 500 random parameter guesses as starting values. Monte Carlo sampling (N = 20,000) was used to generate the model-predicted distributions of rate estimates given by Equations 4, 5, 6, 7, 8, 9, 10, and 11. We obtained discrete categorization probabilities by binning the continuous distributions to the closest stimulus rate among the 4 levels used in the experiment (similar to Körding et al., 2007; Rohe and Noppeney, 2015b).

Analysis of Behavioral Data – Model Comparison

We used Bayesian random-effects model comparison to determine the model that best explains the data at the group level using both Schwarz's Bayesian Information Criterion (BIC) and the corrected Akaike Information Criterion (AICc) (Rigoux et al., 2014). BIC = $-2LL + k \times \ln(n)$, AICc = $-2LL + 2k + 2k(k+1)/(n-k-1)$, corrected for sample size (Cavanaugh, 1997); where LL denotes the log-likelihood, k the number of free parameters, n the total number of data points, and ln the natural logarithm. Log model evidence was obtained for each participant and model by multiplying the BIC or AICc by $-½$. We then calculated each model's posterior frequency and protected exceedance probability (i.e., the probability corrected for chance level that a model is more likely than any others in describing data) using the variational Bayesian analysis (VBA) toolbox (Daunizeau et al., 2014; summarized in Table S1). We also report the models' goodness-of-fit using the generalized coefficient of determination R^2 (Nagelkerke, 1991):

$$R^2=\left\{1-e^{\left(-\frac{2}{n}\left\{LL\left(\hat{\beta}\right)-LL(0)\right\}\right)}\right\}\Big/R^2_{max} \quad \text{(Equation 16)}$$

where $LL(\hat{\beta})$ and $LL(0)$ denote the log-likelihood of the fitted and a 'null' model respectively. The null model describes a chance-level observer with response probability = 0.25 over 4 choices. $R^2_{max}=1-e^{\left(-\frac{2}{n}\{LL(0)\}\right)}$ is a scaling factor proposed by Nagelkerke (1991).

Analysis of behavioral data – Sensory noise function

We considered that the sensory noise (σ^2) in the above sensory likelihood functions might depend on the rate, r, in a power-law fashion (Nieder and Miller, 2003). To determine the precise nature of this dependency, we compared different functions describing the rate-dependent noise using the data from unisensory trials. We started from the following general form:

$$\sigma_r^2=\gamma+p\cdot r^k \quad \text{(Equation 17)}$$

with γ denoting a baseline (i.e., rate independent) noise, r being the rate, p and k, a scaling factor and a power coefficient, respectively. For practical purposes we re-parameterized the power function as follows [here r_1 (σ_1) and r_4 (σ_4) denote the lowest and highest rates (SD of noise), respectively]:

$$\sigma_r^2=\sigma_1^2+\left(r^k-r_1^k\right)\left(\sigma_4^2-\sigma_1^2\right)/\left(r_4^k-r_1^k\right) \quad \text{(Equation 18)}$$

and specifically considered four candidate models describing how parameters change with modality and reliability:

(**Model 1**): parameters being both modality-specific and reliability-specific;
(**Model 2**): parameters being modality-specific but reliability-independent;
(**Model 3**): parameters being both modality-independent and reliability-independent;
(**Model 4**): constant noise across rates, but being both modality-specific and reliability-specific.

We compared these models in capturing the rate dependency of noise using cross-validation on unisensory trials (i.e., partitioning the 22 runs in 5 folds with alternating runs across folds). Model 2 outperformed the others, as demonstrated by an exceedance probability P_{exc} = 0.98 (protected P_{exc} = 0.53) and a mean posterior model probability = 0.47 (SD = 0.056; see Figure S6 for model details).

Preprocessing of MEG Data

All analyses were carried out in MATLAB using SPM12 (Wellcome Trust, London), Fieldtrip (Oostenveld et al., 2011) and custom code. Signal preprocessing was initially carried out on unsegmented MEG data from each run. Infrequent SQUID jumps (observed in 2.3% of the channels, on average) were repaired using piecewise cubic polynomial interpolation. For each participant, we then removed those channels that consistently deviated from the median spectrum (shared variance < 25%) on at least 25% of the runs (number of removed channels = 8.4 on average; SD = 2.2). Environmental magnetic noise was removed using regression based on principal components of reference channels. Both the MEG and reference data were then filtered using a forward-reverse 70 Hz FIR low-pass (−40 dB at 72.5 Hz), a 0.2 Hz elliptic high-pass (−40 dB at 0.1 Hz) and a 50 Hz FIR notch filter (−40 dB at 50 ± 1Hz), and were subsequently re-sampled to 150 Hz. Residual magnetic noise was once more removed using the same regression approach.

ECG and EOG artifacts were removed using independent component analysis ("runica" in Fieldtrip, 30 components) and were identified based on their time course and topographies (Hipp and Siegel, 2013).

MEG data from each run were then segmented into trials. Given the condition-wise differences in reaction times (Figure S1), we aligned the MEG trial-wise segmentations not only to stimulus onset (stimulus-locked window = −0.1 to 0.7 s from stimulus onset) but also to trial-by-trial response onset (response-locked window = −0.7 to 0.1 s from response onset). Segmented MEG data were then corrected on a block-by-block basis to avoid motor contamination of the response on the brain activities specific to each condition, which are our primary interests in subsequent analyses. To this purpose, and independently for each run, the motor-related signals were approximated by averaging MEG data across trials of different experimental conditions but having the same button press (N of conditions = 76, ensuring the same finger response could occur in many distinct conditions) and were finally subtracted from the single-trial MEG data.

Reconstruction of MEG sources

For MEG source analysis, we prepared for each participant a native-space grid of 3.5-mm resolution by re-sampling a group-level anatomical template in native space. The group anatomical template was based on Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL, Ashburner, 2007). A group-level mask was then created by considering non-cerebellum template voxels associated with a gray-matter probability > 0.1, and was back-deformed to the native space of each individual. The native space grid for each individual finally considered 6-connected voxels associated with a participant-specific gray-matter probability > 0.25. Depth-normalized lead fields for each participant were then computed based on a single-shell conductor model and the source-projection filters were derived for each block using a linearly-constrained-minimum-variance (LCMV) beamformer (regularisation = 5%). Data were projected onto the dipole orientation of maximum variance across runs.

Representational similarity analysis

Representational similarity analysis (RSA) quantifies the statistical association between the pairwise dissimilarity of multivariate brain activity in different conditions (MEG representational dissimilarity matrix RDM) and model RDMs that quantify hypotheses about the nature of the neural representations (here the computational models; Kriegeskorte et al., 2008; Walther et al., 2016). We implemented a whole-brain MEG RSA (Giordano et al., 2018) to assess the local encoding of four different types of rate estimates in the MEG activity as predicted by the candidate models fit to behavioral data. The MEG RDMs were computed in native source space of each participant within a spatiotemporal searchlight (10-mm spatial radius, 80-ms temporal extent, and 40-ms temporal overlap between consecutive searchlights). Importantly, we retained the temporal structure within each searchlight, rather than averaging the MEG signal within it. Searchlight-specific RDMs quantified the cross-validated Mahalanobis distances between condition-specific spatiotemporal response patterns (Walther et al., 2016). More specifically, we: **1.** partitioned the 22 runs into 5 folds (alternating consecutive runs across folds); **2.** whitened within-searchlight data for each fold independently using the residuals of a GLM predicting condition-specific response patterns; and **3.** computed the cross-validated Euclidean distance of condition-specific whitened data based on the covariance between pairs of cross-validation folds (average of covariance between the 10 possible pairs of the 5 folds). To construct model RDMs, we first obtained the model-predicted rate estimates for each condition and participant. For each condition, the model-predicted rate was obtained by averaging 20,000 posterior rate estimates (Equations 4, 5, and 10; see above) derived using the best-fitting model parameters. The model RDMs were then constructed by taking the pairwise absolute difference between the model-predicted rate estimates for each pair of the 64 conditions (equivalent to the Euclidean distance between condition-specific rate estimates).

We used the Spearman rank correlation between MEG and model RDM to quantify the MEG representation of the model-predicted rate estimates. The RSA correlation maps were computed in native space, Fisher Z transformed, were then transformed to the DARTEL group space (Gaussian smoothing FWHM = 8 mm), and finally assembled in T statistics for group-level inference. Importantly, the rate predictions from the different models and the associated model RDMs were not independent of each other. As such, a significant RSA correlation with a given model RDM is itself not a strong proof of selective representation because it might be the by-product of a correlated model. To address the key question of selective model representation, we used variance partitioning (see e.g., Hebart et al., 2018; Seibold and McPhee, 1979) to assess the neural encoding of the unique variance ("selective encoding," hereafter) of each model after partialling out all other model RDMs using rank regression (semi-partial correlation). For fusion we partialled out the CI RDM but not the SV and SA RDMs, because fusion by nature is a linear combination of SA and SV, while CI also builds on either SA or SV depending on the task, and hence partialling out all other models would be overly stringent in comparison to the selective encoding tests for the other models (group averaged percent variance retained in the CI, fusion, SA and SV RDMs after partialling out all other model RDMs = 73.7, 80.9%, 75.7 and 80.0%, respectively; SEM $\leq$ 4%). Note that a focus on semi-partial correlations for establishing selective model encoding does not automatically lead to the observation of segregated networks for the encoding of different models, as multiple predictors can exhibit significant partial correlations (see e.g., spatiotemporal overlap of selective encoding of SA and SV models in temporal cortex, and overlap of SA and CI model encoding in cingulate cortex in Figure 3A).

Significance testing relied on a permutation-based random-effects (RFX) framework (shuffling condition labels; 10,000 permutations for each test), in conjunction with a spatiotemporal cluster-mass enhancement of the group-level statistics (parametric cluster-forming threshold of T(14) = 1.76 for 1-sided inference, Maris and Oostenveld, 2007). Correction for multiple comparisons relied, for

each model RDM independently, on the maximum-statistics across the entire analysis space (see below; FWE = 0.05; 1-sided inference for both correlations and semi-partial correlations). Importantly, while correlations were tested across all RSA time windows and gray-matter DARTEL-space voxels, semi-partial correlations were tested only within a mask of significant correlations. In other words, we adopted a two-step approach that initially assessed the non-selective encoding of a particular model, and subsequently tested for selective encoding (Giordano et al., 2018), which mitigated false-positives in a strict way. This RSA was carried out twice, once using the data aligned to stimulus onset, and once aligned to the behavioral response onset. Region-of-interests (ROIs) were identified as the DARTEL coordinates of local MEG model-encoding peaks on the spatiotemporal T(s.p.r) maps and are reported using MNI coordinates in Table 1.

We carried out control analyses to rule out alternative explanations for the coexisting representations of fused and inference representations. The significant RSA effects reported in the main text were contributed by the majority of the participants (Figure S2), and across these the two effects were uncorrelated (FWE-corrected $p > 0.73$ for stimulus-locked ROIs, and $p > 0.66$ for response-locked ROIs; left-sided maximum-statistics permutation tests), which would not be the case if some participants' MEG activity was explained by fusion and that of others by causal inference. Together with the fact that none of the participants' behavior was better described by fusion than CI (Table S1), this suggests that the coexistence of distinct neural representations is unlikely just a by-product of pooling neural data from different participants whose behaviors were best fit by either model. We also performed a control analysis to rule out the possibility that the ROIs with RSA fusion effects simply reflect a fixed linear integration without weighting each modality by its relative reliability. We compared reliability-weighted fusion against a simpler model of fusion that ignores the trial-by-trial variations in auditory reliability. The average P_{exc} of the full reliability-weighted fusion model was as high as 0.875 (SEM = 0.048) across the fusion ROIs.

A further analysis qualified the relative temporal order of the cerebral encoding of the candidate models. We focused on the set of ROIs revealed by the RSA (Table 1), each associated with the selective encoding of a particular model. For each ROI, we implemented a finer-grained RSA (temporal overlap between consecutive searchlights = 6.67 ms, i.e., 1 sample shift for the 150 Hz preprocessed MEG signal) and derived the full time course of the T(s.p.r) statistics measuring the group-level selective model encoding, and then extracted the latencies of the peak T(s.p.r) quantifying the most robust selective encoding of each model. We then used a bootstrap approach (and reporting 95% bias-corrected and accelerated [BCa] bootstrap confidence intervals; Efron and Tibshirani, 1994) to contrast the latency of the peak T for each pair of ROIs encoding the respective models (fusion ROIs versus CI ROIs; Figure S3C–S3F).

Analysis of the neural representations in regions of interest (ROIs)

We implemented a GLM to provide an in-depth understanding of the computational properties of the representations in the ROIs unveiled by the RSA. We focused on the variance of the brain RDMs reflecting the representation of the various candidate computational models. These computation-diagnostic brain RDMs were derived for each ROI (the closest grid point to the ROI's MNI coordinate deformed to native space) within a cross-validated rank regression framework that prevented over-fitting while not favoring *a priori* any particular computational model:

$$\text{MEG RDM} = \beta_0 + \beta_1(\text{CI RDM}) + \beta_2(\text{FU RDM}) + \beta_3(\text{SA RDM}) + \beta_4(\text{SV RDM}) \quad \text{(Equation 19)}$$

where the model RDMs: causal inference = CI, fusion = FU, segregation auditory = SA, segregation visual = SV. We used a leave-one-participant-out cross-validation scheme for estimating the computation-diagnostic RDMs specific to each individual. That is, we derived the betas for predicting in one single GLM the MEG RDMs of all but one participant (independent ranking of data from each participant), and then used these betas to estimate the computation-diagnostic MEG RDM of the left-out participant (using this participant's own model RDMs).

An initial analysis aimed to characterize the dependency of a measure of crossmodal bias derived from the computation-diagnostic MEG RDMs on two key features of flexible multisensory integration: **1.** the interaction of task relevance with sensory reliability, and **2.** the quadratic effect of disparity. The modeling approach mirrored the analysis of behavioral data (Figure 2G) to predict bias within a rank GLM using, as regressors, task, auditory reliability, the task × reliability interaction, disparity and the squared disparity (Equation 1). Here, the cerebral measures of bias and disparity were extracted from the MEG RDMs. The bias was derived as the representational distance (i.e., relevant MEG RDM cells) between each of the 64 conditions on one hand (e.g., a condition with auditory rate = 9.1 Hz and visual rate = 16.4 Hz in auditory task), and the congruent condition with the same task-relevant rate on the other hand (in this example: 9.1 Hz for both auditory and visual rates; still in auditory task). The disparity for each of the 64 conditions was derived as the representational distance between the two respective congruent conditions with the same task-relevant versus task-irrelevant rate (in this example: the two congruent conditions with rates of 9.1 and 16.4 Hz respectively; still in auditory task). Permutation-based RFX inference was used for significance testing with correction for multiple comparisons separately for stimulus- and response-locked ROIs (FWE = 0.05; Figures 4A and S4).

We also visualized the representational geometries in each ROI. We used inter-individual difference multidimensional scaling (INDSCAL; de Leeuw and Mair, 2009; Ashby et al., 1994; Figures 4C, S5A, and S5B) to avoid a known distortion of multidimensional-scaling solutions caused by averaging distances across individuals (Ashby et al., 1994). Here we used INDSCAL models

with different dimensionalities to visualize the influence of different experimental factors upon the neural representations: two-dimensional models for the influence of auditory and visual rates and of their discrepancy, and three-dimensional models for the influence of task and sensory reliability.

Analysis of cerebral representation of categorization behavior

To investigate in which ROIs the MEG activity directly drives participants' categorization behavior, rather than only reflecting the representation of a specific model, we conducted the following analysis. We implemented an RSA measuring the Spearman rank correlation between participant-specific response-locked MEG RDMs and behavioral RDMs, with the latter comprising pairwise absolute distance between the trial-averaged behavioral responses of different conditions. RFX significance testing followed the same permutation approach as for the model-based RSA to ascertain whether the group-average correlation in each ROI was significantly larger than zero (FWE = 0.05 across response-locked ROIs). We first assessed this neuro-behavioral correlation for the entire RDMs pertaining to all 64 conditions. To identify the neural underpinning of the disparity-dependent adaptive behavior, we also quantified the dependency of this neuro-behavioral correlation on disparity. We focused on the ROIs characterized by a significant overall behavioral relevance in the first step, and quantified the RSA effect of behavior independently for the 24 conditions with small-disparity (absolute disparity = 3.6 Hz; Figure 6) and another 24 conditions with large-disparity (absolute disparity > 3.6 Hz). We then tested for a significant modulatory influence of the disparity on the RSA effect of behavior by permuting independently the row and columns of the small and large disparity RDMs, and contrasting their Fisher Z transformed correlation with the respective portions of the behavioral RDM (large minus small; two-sided RFX inference; FWE = 0.05).

DATA AND SOFTWARE AVAILABILITY

Behavioral data and MEG RDMs for each participant in each ROI (Table 1) is available at the Github address specified in the Key Resources Table. Further information and requests for sources and reagents should be directed to and will be fulfilled by the Lead Contact, Yinan Cao (yinan.cao@psy.ox.ac.uk), Department of Experimental Psychology, University of Oxford.